Biblioteca Digital
F C E N - U B A

BIBLIOTECA CENTRAL LUIS F LELOIR
BIBLIOTECA CENTRAL LELOIR
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA

Tesis Doctoral

Fisicoquímica del agua en la superficie proteica y su protagonismo en la unión de carbohidratos

Gauto, Diego F.

2013

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

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

Cita tipo APA:
Gauto, Diego F.. (2013). Fisicoquímica del agua en la superficie proteica y su protagonismo en la unión de carbohidratos. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:
Gauto, Diego F.. "Fisicoquímica del agua en la superficie proteica y su protagonismo en la unión de carbohidratos". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2013.

EXACTAS UBA
Facultad de Ciencias Exactas y Naturales

UBA
Universidad de Buenos Aires

Dirección: Biblioteca Central Dr. Luis F. Leloir, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA - Tel. (++54 +11) 4789-9293

Contacto: digital@bl.fcen.uba.ar

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Inorgánica, Analítica y Química Física

Fisicoquímica del agua en la superficie proteica y su protagonismo en la unión de carbohidratos

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Química Inorgánica, Química Analítica y Química Física

Diego F. Gauto

Director de tesis:
Dr. Marcelo Martí

Consejero de estudios:
Dr. Leonardo Slep

Buenos Aires, 2013

Resumen

La formación de complejos biomoleculares (receptor-ligando), es un proceso que comprende numerosos aspectos físico-químicos, como la formación de diferentes tipos de interacciones entre la proteína y su ligando y el cambio en la estructura de solvatación en las superficies de contacto de ambas moléculas. **Un importante desafío consiste en comprender el papel de las moléculas de agua en el proceso de unión, dado que muchas de estas pueden ser desplazadas desde el sitio de unión hacia el seno del solvente, mientras que otras pueden permanecer fuertemente unidas en el mismo, puenteando la interacción entre la proteína y el ligando**. Por otra parte, el desplazamiento de aguas juega también un rol clave en la determinación del resultado termodinámico del proceso de unión.

El rol de la estructura de solvente en las lectinas, que comprenden todas las proteínas de unión/reconocimiento de carbohidratos sin actividad enzimática, es particularmente relevante en el proceso de reconocimiento, debido a la naturaleza hidrofílica del ligando y el sitio de unión. **En este contexto, el presente trabajo de tesis doctoral, pone a prueba la hipótesis de que la estructura del solvente en el sitio de unión a carbohidratos (CBS) de lectinas, guarda una relación con la forma y las interacciones del complejo lectina-carbohidrato resultante, y por lo tanto, se puede utilizar para predecir su modo de unión**. Para ponerla a prueba, se plantean los siguientes objetivos:

(I) Empleando simulaciones de Dinámica Molecular en solvente explícito, en combinación con un análisis termodinámico estadístico, desarrollar y contribuir a la hipótesis de trabajo de que la estructura formada por las moléculas de agua presentes en el sitio de reconocimiento de carbohidratos en el receptor libre, emulan la posición de los grupos funcionales polares del ligando (en particular los hidroxilos u OHs) cuando se forma el complejo.

(II) Combinando el análisis mencionado en el ítem **I)** con métodos de punto final para el cálculo de la energía libre proteína-ligando, se estudió la base estructural subyacente a la selectividad epimérica mostrada por lectina del hongo *Agaricus bisporus* (ABL), la cual muestra dos sitios de unión a carbohidratos bien diferenciados (CBS-A y CBS-B) que se encuentran en el mismo dominio, en donde la evidencia experimental muestra que cada uno de ellos es capaz de discriminar entre los epímeros, N-Acetil-D-Galactosamina y N-Acetil-D-Glucosamina.

(III) Finalmente, utilizando la información de la estructura de solvente, se ha desarrollado un esquema de *Docking* Molecular modificado que proporciona una mejora significativa en términos de exactitud (la relación entre el complejo predicho y la estructura de referencia cristalográfica) y precisión (discriminación entre el resultado

verdadero de los falsos positivos) en la predicción de la estructura de complejos lectina-carbohidratos.

Palabras claves: lectinas, carbohidratos, sitios de hidratación, dinámica molecular, docking, termodinámica

Abstract

Protein's surface water molecules and their role in the carbohydrate recognition process

The formation of biomolecular complexes (receptor-ligand), is a process that involves many physicochemical aspects, like the formation of different types of interactions between the protein and its ligand, and the change in the solvation structure between the contact surfaces of both molecules. **Understanding the role of water molecules in the binding process is particularly challenging, since while some water molecules are displaced to the bulk solvent, others remain tightly bound in the binding site, bridging protein ligand interactions**. Moreover, water displacement is also a key player in determining the thermodynamic outcome of the binding process.

Lectins, comprise all carbohydrate binding/recognition proteins that do not have enzymatic activity, and given the hydrophilic nature of these ligands and these binding site, the role of the solvent structure in these protein-ligand recognition process is particularly relevant. **In this context, the present PhD thesis, seeks to test and analyze the hypothesis that the solvent structure in the carbohydrate binding site (CBS) of lectins, mimics the shape and interactions of the resulting lectin-carbohydrate complex, and could thus be used to predict their binding mode**. To achieve this task, we have pursued the following specific aims:

(I) Using explicit solvent molecular dynamics simulations (MD) combined with a statistical thermodynamic analysis, we have analyzed the structure adopted by the water molecules in the CBS of several lectins, and related this information with the position of the polar functional groups of the carbohydrate ligands in the resulting complexes.

(II) We have combined the above mentioned analysis with protein-ligand end-point free energy calculations to analyze the underlying structural basis of the epimeric selectivity showed by Agaricus bisporus Lectin (ABL), where two carbohydrate binding site (CBS-A and CBS-B) have found in the same carbohydrate recognition domain (CRD), show that each of them is capable to discriminate between the epimers, N-acetyl-D- Galactosamine and N-acetyl-D-Glucosamine.

(III) Finally, and using the information of the solvent structure, we have devised a modified molecular docking scheme that provides significant improvement in terms of accuracy (the relationship between the predicted and crystallographic complex of reference) and precision (discrimination between the true from false positives results) in the prediction of lectin-carbohydrate complex structures.

Palabras claves: lectins, carbohydrates, water sites, molecular dynamics, docking, termodynamics

A Bautista y Cynthia, mis dos grandes amores

Índice general

Resumen III

Abstract V

1. Introducción **1**
1.1. Introducción General . . . 1
1.1.1. Importancia de la interacción proteína-ligando . . . 3
1.1.2. El problema del agua . . . 6
1.1.3. Sitios de hidratación (WS) . . . 7
1.1.4. Los métodos computacionales y el agua en la superficie proteica . . . 8
1.2. Objetivos . . . 9
1.3. Motivaciones personales . . . 10
Referencias . . . 11

2. Métodos **15**
2.1. Dinámica Molecular (DM) . . . 15
2.1.1. Métodos para explorar la superficie de energía potencial (SEP) . . . 19
2.1.2. Minimización de la Energía . . . 19
2.1.3. Dinámica Molecular (DM) . . . 20
2.1.4. Entorno de Simulación: Condiciones Periódicas de Contorno . . . 21
2.1.5. Condiciones Experimentales: Temperatura y Presión . . . 21
2.1.6. Modelos de agua explícita . . . 22
2.1.7. Validez del campo de fuerzas de AMBER en biomoléculas . . . 23
2.1.8. Campo de fuerza para Carbohidratos . . . 24
2.2. Sitios de hidratación (WS) . . . 25
2.2.1. Identificación de los Sitios de hidratación (WS) . . . 26
2.2.2. Cálculo de las propiedades termodinámicas y dinámicas de los sitios de hidratación (WS) . . . 28
2.3. Método de clusterización para automatizar el análisis estructural y termodinámico de los WS . . . 31
2.3.1. Sitios de hidratación (WS) a partir de la estrategia de clusterización 31
2.3.2. Caracterización termodinámica de los sitios de hidratación (WS) . . 33
2.4. Docking . . . 35
2.4.1. Introducción general a los esquemas de docking utilizados . . . 36
2.4.2. Protocolo y método de docking Sesgado (WSBDM) . . . 39
2.5. Energía Libre de unión . . . 40

2.5.1. Métodos de punto final 42
2.5.2. Métodos clásicos con solvente implícito: MM-PB(SA) y MM-GB(SA) 42
Referencias 45

3. Sitios de hidratación en la superficie de reconocimiento de carbohidratos 53
3.1. Introducción 53
3.2. Ideas e hipótesis de trabajo 54
3.3. Resultados y Discusión 55
3.3.1. Caracterización estructural y termodinámica de los WS en el CBS de CBM32 55
3.3.2. Caracterización estructural y termodinámica de los WS en el CBS de CBM40 56
3.3.3. Caracterización estructural y termodinámica de los WS en el CBS de Gal-3 58
3.3.4. Caracterización estructural y termodinámica de los WS en el CBS de Con-A 59
3.3.5. Caracterización estructural y termodinámica de los WS en el CBS de Cyp-A 61
3.4. Discusión 62
3.5. Conclusión 64
3.6. Metodología específica del capítulo 64
Referencias 65

4. Rol del solvente en la selectividad epimérica 67
4.1. Introducción 67
4.2. Ideas e hipótesis 69
4.3. Resultados 69
4.3.1. Análisis estructural y dinámico del CBS-A de ABL 69
4.3.2. Análisis estructural y dinámico del CBS-B de ABL 72
4.3.3. Análisis de la estructura del solvatación sobre CBS-A de ABL 74
4.3.4. Análisis de la estructura del solvatación sobre CBS-B de ABL 79
4.4. Discusión 82
4.5. Conclusión 85
4.6. Metodología específica del capítulo 85
Referencias 86

5. Mejora de los métodos de predicción de la estructura lectina-carbohidrato 89
5.1. Introducción 89
5.2. Ideas e hipótesis 91
5.3. Resultados 91
5.3.1. Docking de monosacáridos sobre los módulos CBM32 y CBM40 de la Sialidasa Bacteriana de *Clostridium Perfringens* 92
5.3.2. Galectina 3 94
5.3.3. Docking de Galactosa (Gal) sobre Gal-3 96
5.3.4. Docking de N-Acetil-Glucosamina (GlcNAc) sobre Gal-3 97
5.3.5. Docking de N-Acetil-Lactosamina (LAcNAc) sobre Gal-3 99

5.3.6. Docking de trimanósido sobre Concanavalina A (Con-A) 103
5.3.7. Docking de maltotriosa sobre la proteína surfactante D (SP-D) 105
5.3.8. Docking de Hialuronato (HA_4) sobre CD44 107
5.4. Discusión 108
5.4.1. Análisis general de los resultados obtenidos 108
5.4.2. Efecto del tamaño del ligando 112
5.4.3. Efecto del número y característica de los WS utilizados 112
5.4.4. Cómo utilizar el WSBDM 113
5.5. Metodología específica del capítulo 114
Referencias 116

6. Discusión General y Conclusión 119
6.0.1. Perspectivas Futuras 123
6.0.2. Extensión a otros sistemas 124
6.1. Conclusión 125
Referencias 126

7. Publicaciones 129

Agradecimientos 131

Índice de figuras

1.1. Dogma de la biología celular 1
1.2. Dogma de la biología celular extendido 2
1.3. Representación de la asociación proteína-ligando en agua (Martin et al., Annual Review, 2013) 4

2.1. Representación esquemática de los 4 términos del campo de fuerzas molecular. 17
2.2. Representación esquemática de los ángulos de torsión entre monosacáridos vecinos 24
2.3. Definición del ángulo θ y el diedro ϕ. O_w representa el átomo de oxígeno de la molécula de agua y D/A, el átomo donor/aceptor de puente de hidrógeno respectivamente 27
2.4. A) Función de distribución radial $g^{tr}(r)$ y B) gráfico bidimensional de distribución angular para los átomos de oxígeno de los moléculas de agua cercanas al grupo carboxilato de la cadena lateral de la Asp16 de Con-A . . 28
2.5. Energía de interacción promedio de las moléculas de agua dentro del WS con el resto del solvente (línea roja) , con la proteína (línea verde) y total (línea negra). 29
2.6. En puntos negros se observa la probabilidad relativa de aguas (WFP) vs. la distancia tomada desde el centro masas del sitio de hidratación (WS), perteneciente al WS9 al CBS de Concanavalina-A. La línea punteada azul representa el ajuste de una función mono exponencial 30
2.7. Esquema representativo de la determinación de sitios de hidratación (WS) en base al método de clusterización 32
2.8. Representación esquematica de IFST 34
2.9. Representación de mapas de isoenergéticos para un átomo de prueba de tipo oxígeno, sobre una grilla construida sobre el CBS de Gal-3, para el método de docking convencional (CADM) -izquierda-, y modificado (WSBD) - derecha-. 41
2.10. Ciclo termodinámico para el proceso de unión del ligando (L) sobre un dado receptor (R) dando como resultado el complejo (C). El ciclo permite separar los distintos eventos del proceso de unión y asociarlos con la energía libre de unión en vacío, ($\Delta G_{u,vacio}$), panel inferior, de la energía de solvatación en cada subsistema R, L y C ($\Delta G_{R/L/C,solv.}$) 43

3.1. CBM32 con los WSs descriptos en la tabla 3.1 en esferas amarillas A) en la proteína libre y B) en el complejo con galactosa 56

3.2. CBM40 con los WS descriptos en la tabla 3.2 en esferas amarillas A) en la proteína libre y B) en el complejo con ácido siálico 57
3.3. Gal-3 con los sitios de hidratación (WS) descriptos en la tabla 3.3 en esferas amarillas A) en la proteína libre y B) en el complejo con N-Acetil-Lactosamina 59
3.4. Con-A con los sitios de hidratación (WS) descriptos en la tabla 3.4 en esferas amarillas A) en la proteína libre y B) en el complejo con el trimanósido. . . 60
3.5. Ciclofilina A con los WS descriptos en la tabla 3.5 en esferas amarillas A) en la proteína libre y B) en el complejo 62

4.1. Representación esquemática de **A)** β-N-Acetil-D-Glucosamina (GlcNAc) y **B)** β-N-Acetil-D-Galactosamina (GalNAc). El átomo de carbono que está unido al O4 ecuatorial o axial respectivamente, está marcado con * 68
4.2. Representación del dominio de unión de carbohidratos (CRD) de ABL, donde se muestran los dos sitios de unión a carbohidratos (CBS-A y CBS-B) unidos a β-N-Acetil-D-Galactosamina (GalNAc) y β-N-Acetil-D-Glucosamina (GlcNAc), respectivamente. 68
4.3. Estructura representativa de **A)** β-N-Acetil-D-Galactosamina (GalNAc) y **B)** β-N-Acetil-D-Glucosamina (GlcNAc) unidos al CBS-A de ABL 70
4.4. RMSD de los átomos de carbono de los carbohidratos (GalNAc en rojo y GlcNAc en negro) en el CBS-A vs. el tiempo de simulación. Los diferentes estados tautoméricos de las His72 fueron probados: **A)** HIE (histidina protonada en el Nϵ) y **B)** HID (histidina protonada en el Nδ). 71
4.5. Estructura representativa de **A)** β-N-Acetil-D-Glucosamina (GlcNAc) y **B)** β-N-Acetil-D-Galactosamina (GalNAc) unidos al CBS-A de ABL 73
4.6. **A)** Gráfico de g(r) vs. la distancia entre el Gly49O y los átomos de oxígeno del solvente. **B)** Gráfico de distribución bidimensional para los átomos de oxígeno del solvente en que se encuentran en la primera capa de solvatación de Gly49O. **C)** WS sobre el CBS-A en torno a la región que define la especificad epimérica. 75
4.7. **A)** Gráfico de g(r) vs. la distancia entre el Gly49O y el OH4 GalNAc en el CBS-A. **B)** Gráfico de g(r) vs. la distancia entre el Gly49O y el OH4 GlcNAc en el CBS-A. 77
4.8. **A)** Gráfico de distribución angular bidimensional del oxígeno axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Gly49 O. **B)** Gráfico Bidimensional del oxígeno ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Gly49 O. Las regiones delimitadas por las líneas discontinuas rojas representan la posición de los WS, tal cómo se muestra en la figura 4.6 B. 77
4.9. Gráfico de g(r) vs. la distancia de los átomos de oxígeno de las moléculas de agua del solvente con respecto al Nδ de la HIS72. **A)** Para el CBS-A sin ligando de la proteína ABL. **B)** Para el CBS-A unido a su ligando natural, GalNAc. C) Ídem B, pero en el complejo con GlcNAc unido al CBS-A 78
4.10. Gráfico de distribución angular bidimensional del oxígeno de las moléculas de agua del solvente con respecto al Nδ de la HIS72. **A)** Primera capa de solvatación en la proteína libre. **B)** Primera capa de solvatación en el complejo con β-GalNAc (es el observado experimentalmente). C) Ídem B, pero en el complejo con β-GlcNAc 78

4.11. **A)** Gráfico de g(r) vs. la distancia entre el Asp79 CG y los átomos de O de las moléculas de agua del solvente. **B)** Gráfico de distribución angular bidimensional para los átomos de O del solvente en que se encuentran en la primera capa de solvatación de Asp79 CG. **C)** WS sobre el CBS-B en torno a la región que define la especificidad epimérica. 80
4.12. **A)** Gráfico de distribución radial, g(r) del O4 ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Asp 79 CG. **B)** Gráfico de distribución radial, g(r), del O4 axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Asp79 CG. 81
4.13. **A)** Gráfico de distribución angular bidimensional del O4 ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Asp 79 CG. **B)** Gráfico de distribución angular bidimensional del O4 axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Asp79 CG. Las regiones delimitadas por las líneas discontinuas rojas representan la posición de los WS, tal cómo se muestra en la figura 4.11 B. 81

5.1. Esquema de los carbohidratos utilizados en el presente capítulo. Los pares proteína-ligando son: A) CBM32-Gal; B) CBM40-ácido siálico; C) Gal3-LAcNAc; E) ConA-trimanosido; D) CD44-HA_4; F) SPD-Maltotriosa 90
5.2. Cálculo de docking Gal sobre CBM32, donde se compara el CADM con el WSBDM 95
5.3. Cálculo de docking ácido siálico sobre CBM40, donde se compara el CADM con el WSBDM 96
5.4. Cálculo de docking β-D-Gal sobre Gal-3, donde se compara el CADM con el WSBDM 98
5.5. Cálculo de docking GlcNAc sobre Gal-3, donde se compara el CADM con el WSBDM 99
5.6. Cálculo de docking N-acetil-lactosamina sobre Gal-3, donde se compara el CADM con el WSBDM 101
5.7. Cálculo de docking N-acetil-lactosamina sobre Gal-3, con el WSBDM en donde se varia el número de WS utilizados para modificar la grilla 102
5.8. Cálculo de docking de un trimanósido sobre Concanavalina A, donde se compara el CADM con el WSBDM 104
5.9. Cálculo de docking de un trimanósido sobre Concanavalina A con el WSBDM en donde se varia el número de WS utilizados para modificar la grilla 106
5.10. Cálculo de docking de hialuronato sobre CD44, donde se compara el CADM con el WSBDM 109
5.11. Resultados globales comparativos para el CADM y el WSBDM 110

6.1. Gráficos de correlación entre <E>vs. -RTln(WFP) y -TS^e vs. -RTln(WFP) para los WS 121

Índice de Tablas

3.1. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de CBM32 56
3.2. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de CBM40 57
3.3. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de carbohidratos (CBS) de Gal-3 58
3.4. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de Concanavalina A 61
3.5. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de ligandos de CiclofilinaA 62
3.6. Promedio, desvió estándar (s), valores mínimo y máximo para todos los parámetros termodinámicos y estructurales calculados 63

4.1. Porcentajes de ocupancia de puentes de hidrógeno del CBS-A 72
4.2. Valores medios para la energía de interacción calculada utilizando MM-GB(SA) 72
4.3. Porcentajes de ocupancia de puentes de hidrógeno del CBS-B 73
4.4. Valores medios para la energía de interacción calculada con MM-GB(SA) . . 74
4.5. Valores medios de energía de interacción, WFP y R_{min} para los WSs previamente definidos 75
4.6. Valores medios de la energía de interacción calculada con MM-GB(SA) considerando la contribución específica de la molécula de agua en el WS2 . 79
4.7. Valores de energía de interacción, WFP y R_{min} para los WSs previamente definidos 80

5.1. Propiedades de los WS de los CBS para cada proteína utilizada 93

Capítulo 1
Introducción

1.1. Introducción General

El paradigma central de la biología molecular moderna establece que el flujo de información en las células va desde el ADN hacia el ARN y de este último a proteínas. El impacto de este concepto tiene su fuerza en la habilidad que provee para manipular una clase de moléculas basado en el conocimiento de otras o bien, predecir la función de una de ellas en base su semejanza estructural.

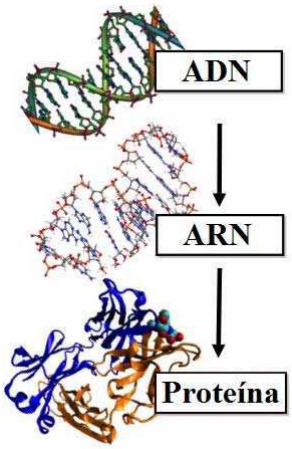

Figura 1.1. Dogma de la biología celular

Pero en realidad, crear una célula requiere también de otras dos clases de moléculas como los lípidos y los carbohidratos. Estas moléculas pueden servir como intermediarios en la generación de energía, como señalizadoras o bien como componentes estructurales y es en este punto, en donde el rol estructural de los carbohidratos adquiere importancia en la construcción de órganos multi-celulares complejos y organismos, los cuales requieren de la interacción de células no solo con otras células, sino con la matriz extra-celular.

En muchos casos, estos glicanos se encuentran como entidades independientes en la superficie externa de las células (dado que la mayor parte de los mismos son secretados),

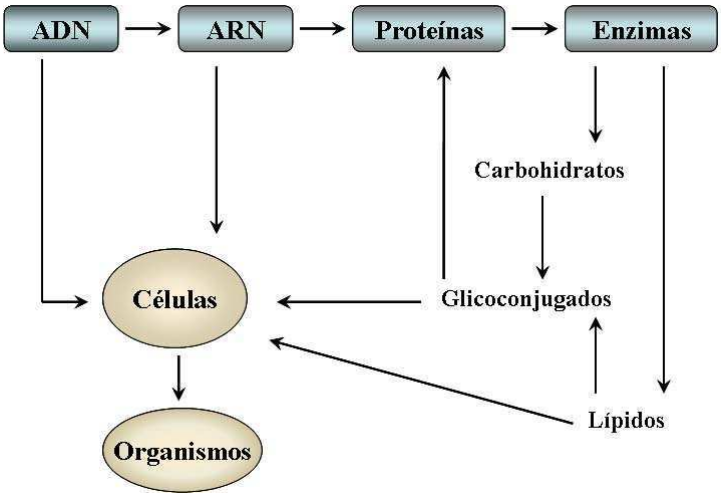

Figura 1.2. Dogma de la biología celular extendido

encontrándose en una posición estratégica para modular o mediar una amplia variedad de eventos vinculados al reconocimiento célula-célula y a las interacciones célula-matriz, las cuales son cruciales para el desarrollo y la función de organismos multicelulares complejos.

Además, en el núcleo y el citoplasma existen numerosos ejemplos de interacción proteína-carbohidrato, donde cumplen la función principal de actuar como llaves regulatorias. Es en este punto en donde resulta interesante plantear un dogma de la biología molecular más global (dogma extendido) y que conlleva a una visión más integradora y menos reduccionista del comportamiento celular, que incluya las interacciones entre proteínas/DNA/RNA-azúcar/lípidos, así como la interacción de los azúcares entre sí y lípidos entre sí dado que juegan un rol no despreciable en modular el comportamiento y el estado celular (ver esquema 1.2).[1]

En las últimas 3 décadas, los progresos en el área de la glicobiología han crecido de manera exponencial, resaltando la importancia de estas macromoléculas en varios procesos intra- e intercelulares. Por un lado, la variabilidad y la complejidad estructural de los glicanos en la superficie celular les permiten funcionar como moléculas señalizadores, de reconocimiento y de adhesión celular. Por ejemplo, los glicanos en la superficie de las células están involucrados en muchas funciones fisiológicas como ser el normal desarrollo embrionario, diferenciación, crecimiento, inhibición por contacto, reconocimiento célula-célula, metástasis, tráfico o localización intracelular, etc. Es así como el campo de la glicobiología ha recibido especial atención y se ha ido acumulando evidencia experimental sobre numerosos tipos de células tumorales, en donde se observa que la transformación a células malignas está fuertemente correlacionada con un cambio en el patrón de glicosilación de las mismas.[2,3]

Centrales en este contexto son las lectinas, que son proteínas de origen no inmunológico

y sin actividad enzimática que se las pude encontrar en la mayoría de los organismos (desde virus y bacterias hasta plantas y animales), responsables del reconocimiento de mono- y oligosacáridos de manera reversible y con alta especificidad.[4–6] Las lectinas no solo son capaces de unirse a oligosacáridos en las células, sino que también pueden reconocer azúcares libres, incluyendo monosacáridos y pese a que la interacción lectina-monosacárido sea relativamente débil, puede ser tan especifica como la interacción antígeno-anticuerpo o enzima sustrato.[5–7] Por otro lado, estudios estructurales llevados a cabo sobre lectinas provenientes de células animales, sugieren que la unión de carbohidratos a lectinas es generada por un número limitado de aminoácidos del receptor, agrupados en lo que se conoce como el sitio de unión de carbohidratos (**CBS** por sus siglas en inglés "Carbohydrate Binding Site"). El CBS por lo general se asocia al extremo no reductor de un oligosacárido unido a glicoproteínas y/o glicolípidos de la membrana celular y este proceso de reconocimiento tiene un fuerte impacto biológico en diversos aspectos vinculados al proceso reconocimiento celular, respuesta inmune e inducción mitogénica, entre otros.[3,8]

1.1.1. Importancia de la interacción proteína-ligando

El reconocimiento molecular en complejos proteína-ligando es responsable de la unión selectiva de ligandos de bajo peso molecular a una proteína específica y de este modo aquellos compuestos que interactúen específicamente con una determinada proteína de interés farmacológico pueden ser utilizados en el tratamiento de diversas enfermedades.

En particular, aquellas enzimas que regulan reacciones bioquímicas son importantes blancos moleculares considerados por la química medicinal. Dos ejemplos representativos de estos podrían ser el inhibidor enzimático trimethoprima, el cual inhibe a la dihidrofolato reductasa bacteriana, que ha sido utilizada en tratamientos contra infecciones bacterianas, mientras que otro ejemplo es la dorzolamida, que ha sido aprobada como droga contra el glaucoma. Si bien, los compuestos de bajo peso molecular pueden interactuar con las macromoléculas a partir de interacciones covalentes (por ejemplo, los trifluorometil ceto-inhibidores de la elastasa, los antibióticos β-lactamicos, la aspirina y otras drogas antiinflamatorias no esteroideas), un amplio espectro de moléculas interactúan a través de interacciones no covalentes proteína-ligando para modular el comportamiento celular.[9]

La mayor contribución a este tipo de interacciones vienen usualmente dadas por numerosas uniones puente de hidrógeno formadas entre la proteína y el ligando. Además, grupos funcionales cargados de manera opuesta se encuentran apareados en el complejo proteína-ligando formando puentes salinos y por otro lado por aquellos grupos lipofílicos del ligando que se localizan en los bolsillos hidrofóbicos del receptor proteico formados por las cadenas laterales no polares de los aminoácidos.

Actualmente, la constante de unión proteína-ligando puede ser estimada por varios

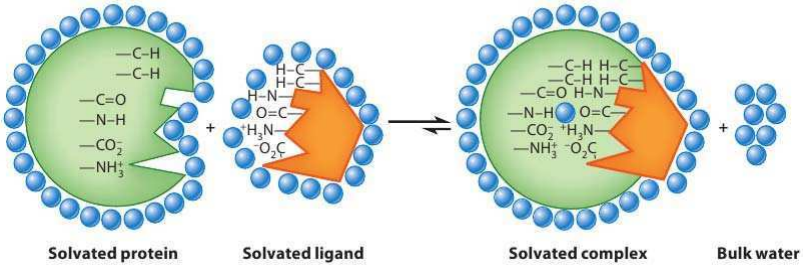

Figura 1.3. Representación de la asociación proteína-ligando en agua (Martin et al., Annual Review, 2013)

métodos experimentales, como por ejemplo mediante Titulación Isotérmica Calorimetrica (ITC), o bien, Resonancia de Plasmones de Superficie (SPR), entre otros. De modo tal que se ha generado un amplio conjunto de datos termodinámicos vinculados al proceso de unión como por ejemplo ΔH_u, ΔS_u y ΔG_u así como constantes de unión, que dan lugar a las siguientes preguntas:[9]

(a) ¿Cuál es la relación de la constante de unión con las propiedades estructurales del complejo?,

(b) ¿Cuál es la propiedad más importante que le permite a una proteína reconocer a un dado ligando y unirse fuertemente al mismo?,

(c) ¿Qué hace que una dada región de la proteína (como por ejemplo un surco en la superficie proteica) sea un sitio de unión a ligandos)?

En resumen, la naturaleza física de la unión proteína-ligando es un fenómeno complejo, del que aún quedan muchas aspectos por comprender. Además, es importante notar que tanto el sitio de unión a ligandos de la proteína como el ligando, se encuentran completamente solvatados previo a la unión y los grupos polares forman uniones puente de hidrógeno con el solvente (ver figura 1.3). Esto genera que en el proceso de unión entre las biomoléculas, se pierda parte de la capa de solvente que rodea la superficie de ambas, que es expelida hacia seno del mismo, involucrando la rotura de numerosos puentes de hidrógeno formados con las moléculas de agua de la primera capa de solvatación.

Por otro lado, es importante tener en cuenta la flexibilidad de las moléculas que participan en el proceso de unión, dado que los ligandos suelen tener varios grados de libertad rotacionales en sus enlaces que en principio, pueden dar lugar a un gran número de conformaciones de baja energía en su estado libre, que se pierden en su estado unido al receptor dadas las restricciones estéricas del sitio de unión. Por otro lado, las proteínas también son flexibles y su conformación en el estado libre puede diferir significativamente del estado

acomplejado, es decir aquel unido al ligando, con una consecuente pérdida en algunos grados de flexibilidad interna de las mismas en el estado unido respecto del estado libre. Como es esperado, ambos efectos traen como consecuencia un costo entrópico asociado a los procesos de unión.[9]

Un caso particular de ligandos, son los carbohidratos, en donde el reconocimiento entre proteína-carbohidrato se establece generalmente a través de redes puente de hidrógeno y contactos complementarios con la superficie proteíca. Es notable que la mayor parte de los residuos no polares de la proteína involucrados en la unión al azúcar son en su mayoría aromáticos, sugiriendo un rol especifico para las interacciones CH-π. Para las lectinas, la mayor parte de las uniones puente de hidrógeno involucran grupos funcionales de las cadenas laterales planos y multivalentes (Asn, Asp, Glu, Gln, Arg, His) y el motivo más frecuente para la unión puente de hidrógeno parece ser NH...OH..O=C. A pesar de que esto incluye numerosas posibilidades, resalta la importancia de la cooperatividad de las uniones puente de hidrógeno dada por los grupos hidroxilo (-OH) de los azúcares. Y si bien la presencia de los grupos -OHs no es frecuente en el sitio de unión de carbohidratos, es común encontrar moléculas de agua en posiciones estratégicas que actúan como puentes entre la proteína y el sustrato. Por otro lado, muchas lectinas requieren la presencia de iones de metales (como por ejemplo Ca^{2+}) para ser completamente activas, aunque no es usual encontrar interacciones directas metal-carbohidrato.[10]

A pesar que el origen de las fuerza involucradas en la unión de los azúcares a proteínas son las mismas que para unir cualquier ligando a un receptor proteico (puentes de hidrógeno, interacciones de vdW y atracciones de carga y dipolares), llama la atención que en general la constante de disociación muestra que la estabilidad del complejo es baja (en el rango de los 100 μM), comparada con otro tipo de ligandos. En general, la baja afinidad esta relacionada con las propiedades intrínsecas de los carbohidratos tales como la falta de cargas y la falta superficies hidrofóbicas extendidas, lo que reduce la probabilidad de formar interacciones fuertes con la proteína. En su lugar, la formación de los complejos proteína-carbohidrato involucran interacciones de vdW débiles y la formación de una red de puentes de hidrógeno entre la proteína y los grupos hidroxilo, acetamida y el átomo de oxígeno del anillo y de la unión glicosídica de los azúcares.[11] Por otro lado, la contribución efectiva de las uniones puente de hidrógeno a la formación del complejo, no depende únicamente de su energía intrínseca, sino además de la energía asociada con el fenómeno de deshidratación y transferencia de moléculas de agua desde el sitio de unión hacia el seno del solvente.[12] En conclusión, es el balance neto de estás contribuciones las que determinan si un puente de hidrógeno es estable o no, independientemente de la relevancia del aducto especifico formado en el proceso de unión. En este contexto, Lemiux et al. han aseverado que la estabilidad de los complejos

proteína-carbohidrato no proviene tanto de la formación de puentes de hidrógeno, sino más bien de la deshidratación de las superficies de contacto.[13]

1.1.2. El problema del agua

Al momento de entender en detalle el proceso de asociación entre biomoléculas, no alcanza con una descripción somera de las contribuciones energéticas entre las mismas, como ser la determinación de la conformación preferencial del ligando o la estimación de las interacciones de corto alcance, por ejemplo, entre los azúcares y los grupos funcionales que interactúan con la proteína. A pesar de que estos aspectos de la interacción biomolecular son muy relevantes, no proveen un abordaje global al proceso de unión entre dos biomoléculas altamente solvatadas como ser los azúcares y las superficie anfifílica del receptor en cuestión. Es decir, si las superficies de ambas biomoléculas son complementarias, las moléculas de agua de ambas superficies deben ser desplazadas, previo a la unión entre las mismas.

En medio acuoso, el receptor de carbohidratos debe expeler aquellas moléculas de agua presentes en el sitio de unión antes de aceptar al sustrato. Sin embargo, un carbohidrato típico se asemeja bastante al conjunto de aguas en el sitio de unión, difiriendo únicamente en la posición precisa de los grupos -OH y la presencia de pequeños parches hidrofóbicos en la superficie. **Esto resalta la importancia de la pre-organización no solo del sitio de reconocimiento, sino de la estructura del solvente aledaña asociada a la naturaleza del reconocimiento entre ambas biomoléculas**.[10,11] Una evidencia experimental de que las moléculas de agua están involucradas en los procesos de asociación molecular entre proteínas y carbohidratos en soluciones acuosas, es la provista por los estudios realizados por Chervenak y Toone. Ellos muestran para un conjunto de proteínas que la cantidad de calor liberado producto de su unión a carbohidratos, se reduce de manera significativa cuando se utiliza D_2O en lugar de H_2O como solvente (esto se explica en que la energía de punto cero para la unión puente de hidrógeno A-D vs. A-H -siendo A un aceptor de puente de hidrógeno, se reduce alrededor de un 10 $\%$ para el A-D). Por otro lado, en base a experimentos basados en espectroscopia Raman, se ha mostrado que la frecuencia libracional es mayor para el H_2O que para el D_2O cuando ambos se encuentran interactuando con la proteína. Esto produce un efecto compensatorio a la contribución del desplazamiento de aguas en el ΔG_u no afectándose la espontaneidad de dicho proceso cuando el experimento se hace en presencia de un solvente o de otro). En general se observa que el ΔH_u difiere entre 1 y 4 Kcal/mol más bajo cuando el experimento se hace en presencia de D_2O con respecto a H_2O, mientras que el cociente entre $\Delta G_u^{H_2O}/\Delta G_u^{D_2O}$ es cercano a la unidad para todos los casos.[14]

Adicionalmente, el análisis de los datos obtenidos a partir de experimentos de titulación isotérmica calorimétrica, sugiere que 25-100 % de la entalpía libre de unión provienen del

proceso de reorganización de las moléculas de agua. Por otro lado, Dunitz et al. mostraron que la liberación de moléculas de agua altamente ordenadas pueden generar un valor de entropía favorable por encima de 2 Kcal/mol a 300 K.[15] Esta estimación está basada en la comparación de calores de hidratación estándar para varias sales inorgánicas anhidras e hidratadas. **En este contexto, es claro que el proceso de unión de ligandos conlleva a una significativa reorganización del solvente en torno a las superficies de contacto, que contribuye de manera significativa a la energía libre de unión**.

Por consiguiente, en las últimas décadas, se ha prestado particular atención a estudiar la estructura y la dinámica de las moléculas de agua en el sitio de unión de ligandos y el rol que estas juegan en mediar el proceso de reconocimiento por parte de las proteínas.[16–19] Como ya se ha planteado, en sus estados libres, tanto las proteínas como los ligandos están hidratados y de hecho, la cristalografía y las simulaciones de dinámica molecular han demostrado que las proteínas suelen tener moléculas discretas de agua en el sitio de unión de carbohidratos, localizadas en posiciones aproximadamente equivalentes en los sitios en los que los grupos hidroxilo serán encontrados en el complejo.[20] **En este contexto, es valido plantearse que si los ligandos son carbohidratos, habrá una pequeña ganancia en la entalpía esperada de reemplazar una agua por un grupo hidroxilo del ligando. En contraste, habrá un cambio significativo en la entropía asociada al proceso de desplazar las moléculas de aguas unidas de cada sitio interactuante**.[21,22]

1.1.3. Sitios de hidratación (WS)

Las moléculas de agua que interaccionan con los residuos polares del sitio activo de una proteína están por lo general desfavorecidas desde un punto de vista entrópico, dadas las restricciones traslacionales y orientacionales que le impone la interacción con el soluto o bien,[15] dependiendo de la hidrofobicidad del sitio de unión pueden estar energéticamente desfavorecidas dado que las moléculas de agua en el sitio de unión no pueden interactuar de manera completa y óptima con su entorno.[23]
En los últimos años se ha prestado una enorme atención al estudio de la estructura y la dinámica de las moléculas de agua en el sitio de unión de proteínas que unen carbohidratos dado que las correspondientes superficies de contacto son altamente hidrofílicas, especialmente si se las compara con aquellas típicamente hidrofóbicas, como ser las interacciones proteína-proteína y proteína-droga. En este contexto, los efectos del desplazamiento de una molécula de agua en el sitio de unión de carbohidratos de Concanavalina A con diferentes trimanósidos a la energía de unión fue evaluado, mostrando que el re-ordenamiento de estás moléculas de agua contribuyen favorablemente a aumentar la afinidad de unión.[19] De este modo, los sitios de hidratación (**WS** por sus siglas en inglés ”Water Sites") se definen como aquellas regiones del espacio cercanas a la superficie proteíca en donde la probabili-

dad de encontrar moléculas de agua sea mayor a la probabilidad de encontrarlas en el seno del solvente en una región del mismo volumen (típicamente 1 $Å^3$). Este incremento en la probabilidad en las regiones aledañas a la superficie proteíca tiene su fundamento en que las moléculas de agua son más retenidas por las regiones polares de la misma que por las moléculas de agua en sí.

Recientemente, Saraboji et al. han reportado la estructura de Galectina 3 (Gal3-C) en su estado libre y unida a lactosa, por medio de experimentos de cristalografia de rayos X y RMN de alta resolución, mostrando no solo la presencia de moléculas de agua fuertemente retenidas en el CBS sino también la similitud existente en la posición de las mismas en el CBS de la proteína en su estado libre y los átomos de oxígeno de los grupos hidroxilo de la lactosa en el complejo proteína-ligando, indicando que el CBS de Gal3-C esta pre-organizado por el solvente para coordinar los átomos de oxígeno del ligando en un arreglo óptimo para el reconocimiento del carbohidrato.[11]

También existe un precedente en el diseño racional de drogas guiado por el desplazamiento de moléculas de agua claves presentes en el sitio de unión de ligando, puede verse en el trabajo de Lam et al., en donde se muestra que el desplazamiento de una molécula de agua localizada en el bolsillo de unión del complejo HIV-1/KNI-272 por un inhibidor similar al KNI-272 hace que este presente mayor afinidad por el receptor que el inhibidor sin modificar.[24] Por consiguiente, la liberación de estás moléculas de agua parecería contribuir favorablemente a la energía libre de unión del nuevo inhibidor. Otros trabajos confirman esta idea a partir de simulaciones de dinámica molecular, donde los autores demuestran que aquellos ligandos diseñados para desplazar moléculas de agua altamente ordenadas, exhiben un aumento en su afinidad[25,26] y por lo tanto, es esperable que el estudio de las propiedades termodinámicas de las mismas pueda dar información sumamente útil, no solo en relación al mecanismo de unión de carbohidratos y su termodinámica sino también como guía para el diseño racional de drogas.

1.1.4. Los métodos computacionales y el agua en la superficie proteica

Las simulaciones de dinámica molecular surgen como una importante herramienta para comprender las bases físicas de la estructura y función de biomoléculas, proveyendo los detalles de los movimientos de las partículas en función del tiempo. De esta manera, esta herramienta puede ser utilizada para abordar preguntas específicas acerca de las propiedades de un sistema modelo, frecuentemente de manera más sencilla que la alcanzada con los experimentos actuales.[27] Por ejemplo, las simulaciones de dinámica molecular en solvente explícito surgen como una promisoria estrategia para el estudio de las interacciones agua-proteína, permitiendo alcanzar una descripción de la estructura y de las propiedades

dinámicas de las moléculas de agua presentes en el CBS. Además, utilizando simulaciones de DM en agua explícita, combinada con un análisis termoestadístico, las propiedades termodinámicas de las moléculas de agua presentes en el sitio de unión pueden ser cálculadas y analizadas desde un punto de vista comparativo.[28,29] Por ejemplo, la energía libre de unión estándar de unir una molécula de agua en el sitio de unión de una proteína ha sido estudiado utilizando el método de doble-desacoplamiento, una metodología que permite el cálculo de las energías involucradas en la transferencia de moléculas de agua desde la superficie de la proteína hacia el seno del solvente.[30–32]

En otro esquema estratégico, Li et al., utilizando la teoría del fluido inhomogéneo, han sido capaces de calcular la contribución entálpica y entrópica a la energía libre de llevar una molécula de agua desde el seno del solvente hacia la superficie proteica para un grupo seleccionado de moléculas de agua presentes en el CBS.[19,29,33–35] Por su parte, Abel et al., en base a los trabajos de Li y Lazaridis, han estudiado la distribución de las moléculas de agua en el sitio de unión de ligandos del factor Xa, un importante blanco para el diseño de drogas involucradas en la vía de la coagulación, con el objetivo de poder diseñar una función semiempírica que discrimine aquellas drogas que se unen de las que no se unen al fXa.[28] Finalmente, Di Lella et al. han estudiado las propiedades de la estructura de solvatación sobre el CBS de Galectina 1 (Gal-1) a partir de simulaciones de DM., mostrando que las moléculas de agua presentes en el CBS juegan un rol activo en la unión proteína-ligando y que es posible predecir las regiones del espacio donde los átomos de oxígeno del ligando pueden ser encontrados en la estructura del complejo Gal-1/Galactosamina.[36]

1.2. Objetivos

El presente trabajo de tesis doctoral, se propone poner a prueba **la hipótesis de que la estructura del solvente en el sitio de unión a carbohidratos (CBS) de lectinas, guarda una relación con la forma y las interacciones del complejo lectina-carbohidrato resultante**, y por lo tanto, se puede utilizar para predecir su modo de unión. Para esto, se plantean los siguientes objetivos:

(I) Emplear simulaciones de Dinámica Molecular en solvente explícito, en combinación con un análisis termodinámico estadístico, para desarrollar y contribuir a la hipótesis de trabajo de que la estructura formada por las moléculas de agua presentes en el sitio de reconocimiento de carbohidratos en el receptor libre, emulan la posición de los grupos funcionales polares del ligando (en particular los hidroxilos u OHs) cuando se forma el complejo.

(II) Combinar el análisis mencionado en el ítem **I)** con métodos de punto final para el cálculo de la energía libre proteína-ligando, con el objeto de comprender la base

estructural subyacente a la selectividad epimérica mostrada por lectina del hongo *A*garicus bisporus (ABL), la cual presenta dos sitios de unión a carbohidratos bien diferenciados (CBS -A y CBS- B) que se encuentran en el mismo dominio, en donde la evidencia experimental muestra que cada uno de ellos es capaz de discriminar entre los epímeros, N-acetil-D-Galactosamina y N-acetil-D-Glucosamina.

(III) Utilizar la información de la estructura de solvente, desarrollando un esquema de Docking Molecular Modificado que proporcione una mejora significativa en términos de exactitud (la relación entre el complejo predicho y la estructura de referencia cristalográfica) y precisión (discriminación entre el resultado verdadero de los falsos positivos) en la predicción de la estructura de complejos lectina-carbohidratos con respecto a los métodos de docking convencionales.

(IV) Finalmente, desarrollar los programas que permitan caracterizar la estructura de solvatación de las moléculas de agua en la superficie proteíca a partir de los resultados provistos por las simulaciones de DM de proteínas en solvente explícito, que puedan ser utilizados por cualquier persona que este interesada en aplicar esta clase de análisis a los sistemas pertinentes a sus proyectos de investigación.

1.3. Motivaciones personales

Las motivaciones personales que me llevaron a abordar este trabajo de tesis estuvieron basadas en poder alcanzar un entendimiento general de la estructura y dinámica de las biomoléculas utilizando las herramientas que surgen de la fisicoquímica. Desde un primer momento, sentí que la comprensión de estos fenómenos desde la ciencia básica son centrales para entender, desde un punto de vista no tan reduccionista, el comportamiento celular y la aplicación al diseño de fármacos y de diversas estrategias para poder modular desde un punto de vista racional el comportamiento de las mismas. De este modo, los métodos computacionales me proveyeron la libertad necesaria para mudarme de un sistema a otro y de poder desarrollar diversas metodologías donde el cuello de botella fuera mi imaginación y no tanto los recursos disponibles, que si bien son centrales, no lo son tanto como para abordar un proyecto experimental. Por último, durante mi formación he priorizado la diversidad de enfoques, desde la bioinformática hasta la termodinámica estadística con el propósito de desarrollar una visión integral que me permitiera atacar determinados interrogantes desde distintos flancos.

Referencias

[1] Ajit Varki, Richard Cummings, and Jeffrey Esko. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, August 2002.

[2] Yoko Funakoshi and Tadashi Suzuki. Glycobiology in the cytosol: the bitter side of a sweet world. *Biochimica et biophysica acta*, 1790(2):81–94, February 2009.

[3] Haike Ghazarian, Brian Idoni, and Steven B Oppenheimer. A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. *Acta histochemica*, 113(3):236–47, May 2011.

[4] A. Nurisso, B. Blanchard, A. Audfray, L. Rydner, S. Oscarson, A. Varrot, and A. Imberty. Role of water molecules in structure and energetics of Pseudomonas aeruginosa lectin I interacting with disaccharides. *Journal of Biological Chemistry*, 285(26):20316–20327, 2010. cited By (since 1996)11.

[5] T. Minko. Drug targeting to the colon with lectins and neoglycoconjugates. *Advanced Drug Delivery Reviews*, 56(4):491–509, 2004. cited By (since 1996)88.

[6] E. Gorelik, U. Galili, and A. Raz. On the role of cell surface carbohydrates and their binding proteins (lectins) in tumor metastasis. *Cancer and Metastasis Reviews*, 20(3-4):245–277, 2001. cited By (since 1996)144.

[7] G.A. Rabinovich, M.A. Toscano, S.S. Jackson, and G.R. Vasta. Functions of cell surface galectin-glycoprotein lattices. *Current Opinion in Structural Biology*, 17(5):513–520, 2007. cited By (since 1996)109.

[8] Halina Lis and Nathan Sharon. Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition. *Chemical reviews*, 98(2):637–674, April 1998.

[9] H.-J. Bohm and G. Klebe. What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angewandte Chemie - International Edition in English*, 35(22):2588–2614, 1996. cited By (since 1996)179.

[10] A.P. Davis and R.S. Wareham. Carbohydrate recognition through noncovalent interactions: A challenge for biomimetic and supramolecular chemistry. *Angewandte Chemie - International Edition*, 38(20):2978–2996, 1999. cited By (since 1996)285.

[11] Kadhirvel Saraboji, Samuel Genheden, Carl Diehl, Johan Qvist, Ulrich Weininger, Ulf J Nilsson, Hakon Leffler, Ulf Ryde, Mikael Akke, Derek T Logan, and Maria Hå kansson. The

Carbohydrate-Binding Site in Galectin-3 Is Preorganized To Recognize a Sugarlike Framework of Oxygens: Ultra-High-Resolution Structures and Water Dynamics. *Biochemistry*, 51(1):296–306, January 2012.

[12] A. Ben-Naim. The role of hydrogen bonds in protein folding and protein association. *Journal of Physical Chemistry*, 95(3):1437–1444, 1991. cited By (since 1996)30.

[13] Raymond U. Lemieux. How Water Provides the Impetus for Molecular Recognition in Aqueous Solution. 4842(4):373–380, 1996.

[14] M.C. Chervenak and E.J. Toone. A direct measure of the contribution of solvent reorganization to the enthalpy of ligand binding. *Journal of the American Chemical Society*, 116(23):10533–10539, 1994. cited By (since 1996)160.

[15] J.D. Dunitz. The entropic cost of bound water in crystals and biomolecules. *Science*, 264(5159):670, 1994. cited By (since 1996) 342.

[16] C. Clarke, R.J. Woods, J. Gluska, A. Cooper, M.A. Nutley, and G.-J. Boons. Involvement of water in carbohydrate-protein binding. *Journal of the American Chemical Society*, 123(49):12238–12247, 2001. cited By (since 1996)98.

[17] R. Kadirvelraj, B.L. Foley, J.D. Dyekjær, and R.J. Woods. Involvement of water in carbohydrate-protein binding: Concanavalin A revisited. *Journal of the American Chemical Society*, 130(50):16933–16942, 2008. cited By (since 1996)33.

[18] S. Di Lella, L. Ma, J.C. Díaz Ricci, G.A. Rabinovich, S.A. Asher, and R.M.S. Álvarez. Critical role of the solvent environment in galectin-1 binding to the disaccharide lactose. *Biochemistry*, 48(4):786–791, 2009. cited By (since 1996)11.

[19] Z. Li and T. Lazaridis. The effect of water displacement on binding thermodynamics: Concanavalin A. *Journal of Physical Chemistry B*, 109(1):662–670, 2005. cited By (since 1996)57.

[20] N.K. Mishra, P. Kulhánek, L. Šnajdrová, M. Petřek, A. Imberty, and J. Koča. Molecular dynamics study of pseudomonas aeruginosa lectin-ii complexed with monosaccharides. *Proteins: Structure, Function and Genetics*, 72(1):382–392, 2008. cited By (since 1996)13.

[21] N. Shimokhina, A. Bronowska, and S.W. Homans. Contribution of ligand desolvation to binding thermodynamics in a ligand-protein interaction. *Angewandte Chemie - International Edition*, 45(38):6374–6376, 2006. cited By (since 1996)16.

[22] B.A. Williams, M.C. Chervenak, and E.J. Toone. Energetics of lectin-carbohydrate binding. a microcalorimetric investigation of Concanavalin A-oligomannoside complexation. *Journal of Biological Chemistry*, 267(32):22907–22911, 1992. cited By (since 1996)68.

[23] Piotr Setny, Riccardo Baron, and J Andrew McCammon. How Can Hydrophobic Association Be Enthalpy Driven? *Journal of chemical theory and computation*, 6(9):2866–2871, September 2010.

[24] P.Y.S. Lam, P.K. Jadhav, C.J. Eyermann, C.N. Hodge, Y. Ru, L.T. Bacheler, J.L. Meek, M.J. Otto, M.M. Rayner, Y.N. Wong, C.-H. Chang, P.C. Weber, D.A. Jackson, T.R. Sharpe, and S. Erickson-Viitanen. Rational design of potent, bioavailable, nonpeptide cyclic ureas as hiv protease inhibitors. *Science*, 263(5145):380–384, 1994. cited By (since 1996) 595.

[25] P.R. Connelly, R.A. Aldape, F.J. Bruzzese, S.P. Chambers, M.J. Fitzgibbon, M.A. Fleming, S. Itoh, D.J. Livingston, M.A. Navia, J.A. Thomson, and K.P. Wilson. Enthalpy of hydrogen bond formation in a protein-ligand binding reaction. *Proceedings of the National Academy of Sciences of the United States of America*, 91(5):1964–1968, 1994. cited By (since 1996) 92.

[26] K.A. Watson, E.P. Mitchell, L.N. Johnson, Jong Chan Son, C.J.F. Bichard, M.G. Orchard, G.W.J. Fleet, N.G. Oikonomakos, D.D. Leonidas, M. Kontou, and A. Papageorgioui. Design of inhibitors of glycogen phosphorylase: A study of α- and β-c- glucosides and 1-thio-β-d-glucose compounds. *Biochemistry*, 33(19):5745–5758, 1994. cited By (since 1996) 95.

[27] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, 2002. cited By (since 1996)754.

[28] Robert Abel, Tom Young, Ramy Farid, Bruce J. Berne, and Richard A. Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008.

[29] Z. Li and T. Lazaridis. Thermodynamics of buried water clusters at a protein-ligand binding interface. *Journal of Physical Chemistry B*, 110(3):1464–1475, 2006. cited By (since 1996) 34.

[30] J. Pranata and W.L. Jorgensen. Monte carlo simulations yield absolute free energies of binding for guanine-cytosine and adenine-uracil base pairs in chloroform. *Tetrahedron*, 47(14-15):2491–2501, 1991. cited By (since 1996) 18.

[31] M.K. Gilson, J.A. Given, B.L. Bush, and J.A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysical Journal*, 72(3):1047–1069, 1997. cited By (since 1996) 457.

[32] D. Hamelberg and J.A. McCammon. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: Double-decoupling method. *Journal of the American Chemical Society*, 126(24):7683–7689, 2004. cited By (since 1996) 102.

[33] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 1. theory. *Journal of Physical Chemistry B*, 102(18):3531–3541, 1998. cited By (since 1996) 52.

[34] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids. *Journal of Physical Chemistry B*, 102(18):3542–3550, 1998. cited By (since 1996) 32.

[35] Z. Li and T. Lazaridis. Thermodynamic contributions of the ordered water molecule in hiv-1 protease. *Journal of the American Chemical Society*, 125(22):6636–6637, 2003. cited By (since 1996) 48.

[36] S.D. Lella, M.A. Martí, R.M.S. Álvarez, D.A. Estrin, and J.C. Díaz Ricci. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *Journal of Physical Chemistry B*, 111(25):7360–7366, 2007. cited By (since 1996) 11.

Capítulo 2
Métodos

2.1. Dinámica Molecular (DM)

La simulación por dinámica molecular clásica (**DM**) de biomoléculas, que comenzó a utilizarse hace aproximadamente 30 años atrás, es actualmente una de las herramientas más utilizadas para investigar la estructura y la dinámica de las proteínas. Entre sus aplicaciones contamos con el estudio de unión de ligandos, la dilucidación de mecanismos de reacción, problemas vinculados con la desnaturalización y plegado proteico, la complementación de resultados experimentales y el refinamiento de estructuras resueltas por resonancia magnética nuclear (RMN) y cristalografía de rayos-X.[1,2] En particular, los métodos de simulación por dinámica molecular permiten obtener información relacionada con las conformaciones de proteínas y otras macromoléculas de interés biológico, así como información cinética y termodinámica de los procesos de interés. De este modo, las simulaciones por DM pueden proveer el detalle fino concerniente a los movimientos de partículas individuales en función del tiempo y esto puede ser utilizado para cuantificar las propiedades de un sistema con un nivel de precisión y en una escala temporal que de otro modo sería inaccesible.

En este contexto, la simulación computacional de biomoléculas surge como una promisoria herramienta para extender nuestra comprensión sobre los sistemas biológicos.[3] Por ejemplo, un área que tiene una fuerte simbiosis entre las simulaciones computacionales y el campo experimental es la resonancia magnética nuclear de proteínas, dado que complementa la capacidad de esta técnica para predecir la estructura y para poder determinar las propiedades termodinámicas y dinámicas de las mismas. Un trabajo reciente, resalta la importancia de los métodos computacionales, mostrando que si bien existe una buena correlación para el cálculo de entropía alcanzado mediante DM y RMN, solamente se tiene en cuenta el 25 % de la entropía total si solo se utiliza RMN.[4,5] Por otro lado, a diferencia de los cálculos mecano-cuánticos, en los que se debe tratar de manera explícita los electrones del sistema (a pesar de que algunos de ellos puedan ser ignorados, como en el caso de los métodos cuánticos semi-empíricos), los métodos que involucran las aproximaciones de la dinámica molecular clásica ignoran los movimientos electrónicos y solamente calculan

la energía del sistema en el marco de la aproximación de Born-Oppenheimer, como una función dependiente de la posición de los núcleos; lo cual claramente reduce el número de operaciones que se deben hacer para dicho cálculo energético en cada configuración del sistema, disminuyendo drásticamente el costo computacional y haciendo posible el estudio de sistemas mucho más grandes, que serían temporalmente inabordables si se intentarán describir a partir de un modelo cuántico.[6] Por otro lado, las simulaciones de DM están basadas en la propagación determinista de las fuerzas que actúan sobre un sistema molecular[7] de acuerdo a las ecuaciones de newton, lo que hace posible observar a un nivel atomístico la progresión del movimiento molecular[8] en una escala de tiempo del orden 1.0 10^{-12} a 1.0 10^{-9} segundos, como por ejemplo, las vibraciones elásticas de los dominios globulares, la rotación de las cadenas laterales de los aminoácidos expuestas a solvente y el movimiento torsional libracional de aquellas moléculas cuya libertad traslacional y rotacional, se encuentra limitada por su interacción con el entorno.[3]

A la función de energía potencial (dependiente de las posiciones nucleares) se la denomina campo de fuerzas, ya que el gradiente de la misma permite obtener la fuerza que actúan sobre cada átomo, aunque también se lo describe en la literatura como potencial. Actualmente, hay disponibles diferentes campos de fuerza clásicos como CHARMM,[9] GROMOS[10] o AMBER.[11] Todos poseen una descripción similar de los términos que contribuyen a la energía, aunque difieren en los parámetros que se utilizan en la ecuaciones. En esta tesis se ha utilizado el campo de fuerzas de AMBER que será descripto con mayor detalle a continuación. En general, la mayor parte de los campos de fuerzas que se utilizan actualmente pueden ser interpretados en función de 4 términos relativamente simples, que incluyen descripciones de las fuerza intra- e intermoleculares con el sistema, los cuales imponen una penalidad energética a las desviaciones de las posiciones en los parámetros de equilibrio (distancia, ángulo, ángulos diedro, etc.). De este modo, una forma funcional que se suele utilizar para modelar un conjunto de átomos y/o moléculas se puede observar en la ecuación 2.1,

$$V(\mathbf{r}^N) = \sum_{enlaces} K_B(l_i - l_{i,0})^2 + \sum_{'angulo} K_\theta(\theta_i - \theta_{i,0})^2 + \sum_{torsiones} \frac{V_n}{2}(1 + cos(n\omega - \gamma))$$
$$+ \sum_{i}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \tag{2.1}$$

en donde $V(\mathbf{r}^N)$ es la energía potencial expresada en función de la posición **r** de N partículas (usualmente átomos). El primer término corresponde al estiramiento de los enlaces entre pares de átomos, en donde la unión entre los mismos es descripta por un oscilador armónico de acuerdo a la ley de Hooke, el cual tiene su mínimo a la distancia de equilibrio $l_{i,0}$, que

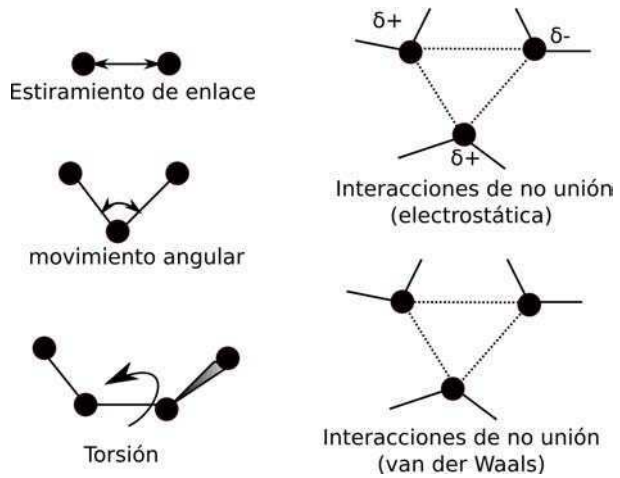

Figura 2.1. Representación esquemática de los 4 términos del campo de fuerzas molecular.

correspondería a la distancia de equilibrio entre esos dos átomos, y K_B es la constante de fuerza del oscilador (si bien la curva de energía potencial de un enlace real no es exactamente armónica, está aproximación resulta razonablemente buena si las distancias de enlace no difieren mucho de su posición de equilibrio). El segundo término da cuenta de la desviación angular respecto de un valor de referencia $\theta_{i,0}$ en el marco de una aproximación armónica con K_θ la constante de fuerza de la flexión angular. Típicamente, los ángulos son más susceptibles a la deformación que los enlaces, de modo que las constantes de fuerza angulares poseen valores de entre 30-90 kcal/(mol Rad) mientras que las de enlace poseen valores de entre 200-500 kcal/(mol Å). El tercer término corresponde a la energía de torsión de los ángulos denominados diedros o impropios, los que describen la energía potencial a lo largo de un eje de rotación definido por un enlace. En la fórmula del campo de fuerzas, ω es el valor que toma el diedro que se calcula, de acuerdo al esquema de la figura 2.2. El término V_n, está relacionado con la magnitud de la barrera entre dos mínimos adyacentes, "n" es la multiplicidad, que describe cuantos mínimos hay para la rotación entre 0 y 360 grados. Por último, "γ" es la fase que determina donde el ángulo pasa por el primer mínimo. Por ejemplo, el perfil de energía para la rotación alrededor de un enlace simple de dos átomos de carbono unidos con hibridación sp^3 podría ser representado por un término rotacional con n=3 (correspondiente a los tres mínimos 60, 180 y 240 grados) y $\gamma = 60^0$ que es el primer mínimo correspondiente a la conformación anti. Por último, los campos de fuerzas incluyen uno o más términos de no unión (últimos términos en la ecuación 2.1), los cuales tienen en cuenta a moléculas independientes y todos los átomos a más de tres enlaces interactúan entre sí a través de fuerzas de términos de no unión, las cuáles juegan un rol primordial en determinar la estructura de especies moleculares individuales. Las interacciones de no unión no dependen de una interacción específica entre los átomos,

sino que más bien son interacciones a través del espacio usualmente modeladas como funciones de la distancia. Los términos de no unión en un campo de fuerzas son usualmente considerados en dos grupos, uno que comprende las interacciones electrostáticas y otro que comprende las interacciones de van der Waals (**vdW**). La interacción electrostática es una consecuencia de la distribución asimétrica de la densidad de carga en las moléculas. En el potencial AMBER, el potencial electrostático de cada molécula es representado por un arreglo de cargas fraccionarias puntuales ubicadas en la posición de los átomos de la molécula. Estas cargas fraccionarias están elegidas de manera tal que reproduzcan el potencial electrostático por la molécula sobre su entorno.[12,13] Por ejemplo, en uno de los modelos de la molécula de agua usado por AMBER, al átomo de oxígeno le es asignado una carga de -0.8 **e-** y a los átomos de hidrógeno, una carga de 0.4 **e-** cada uno.

Una vez asignadas las cargas a todos los átomos, la interacción entre los mismos es calculada usando la ley de Coulomb (ver último término de la ecuación 2.1). No obstante, las interacciones electrostáticas no pueden por sí solas describir totalmente las interacciones entre los átomos de moléculas diferentes. Por ejemplo, en gases nobles donde los momentos multipolares son todos cero, no se esperaría que haya atracción entre los átomos, aunque sin embargo es posible observar la presencia de fases líquidas y sólidas en determinadas condiciones de presión y temperatura. Estás desviaciones a la idealidad del comportamiento de un gas noble fueron tempranamente cuantificadas por vdW y años más tarde, London mostró que esta contribución atractiva es originada por fuerzas dispersivas que pueden ser explicadas en base a la mecánica cuántica a través de la generación de dipolos transientes, dados por la fluctuación de la posición de los electrones en la nube electrónica.

Por otro lado, si solo tuviéramos en cuenta la interacción electrostática de dos átomos con signos opuestos, estos se atraerían hasta colapsar en un mismo punto del espacio obteniéndose un valor infinito de energía, lo cual muestra la incompletitud del modelo coulómbico para describir las interacciones de no unión. Un ejemplo se puede ver en el acercamiento de dos átomos de Argón por debajo de los 3 Å en donde un pequeño decremento de la distancia genera un gran aumento en la energía potencial (este incremento también tiene su origen en la mecánica cuántica y puede ser comprendido en términos del principio de exclusión de Pauli). La fórmula que describe de manera adecuada a este tipo de interacción, repulsiva de corto alcance y dispersiva a largo alcance, es el potencial denominado de Lennard-Jones 12-6, el cual posee dos parámetros que dependen del par de átomos del cual se esta describiendo la interacción: la distancia del mínimo de energía entre ambos átomos (r_{ij}) y el valor de la energía en ese punto (ϵ_{ij}). Los parámetros están tabulados según el tipo de átomo, de manera que para encontrar los parámetros correspondientes a la interacción del par de átomos i y j, se utilizan las siguientes reglas de combinación de Lorentz-Berthelot en donde podemos escribir al diámetro de colisión como

$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj})$; y a la profundidad del pozo como $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$.

El cálculo del término de no unión es el más costoso desde el punto de vista computacional ya que en el caso ideal involucra computar las interacciones de aproximadamente Nx(N-1) pares de átomos, mientras que los términos de enlace, angulares y de torsión escalan linealmente con el número de átomos (N). De este modo, para disminuir el costo computacional, típicamente las interacciones de no unión son cálculadas solamente entre los pares de átomos que se encuentran a una distancia menor que un valor de corte conveniente (usualmente los valores del radio de corte están alrededor de 12 Å).[6]

2.1.1. Métodos para explorar la superficie de energía potencial (SEP)

Para la mayor parte de los casos, a excepción de sistemas muy simples, la energía potencial del sistema requiere ser resuelta por una función multidimensional de las 3N coordenadas (3 coordenadas para cada átomo), lo que la hace difícil de resolver analíticamente. A esta función 3N dimensional que describe la energía en función de la posición de N átomos se la conoce como superficie de energía potencial **(SEP)** y la importancia de conocer la SEP en su completitud radica en que nos permitiría obtener todas las propiedades termodinámicas y cinéticas del sistema. Sin embargo, esta es una tarea inabordable y es por eso que se han desarrollado métodos alternativos para explorar la SEP que nos permiten obtener los valores termodinámicos deseados de manera computacionalmente eficiente.[6]

2.1.2. Minimización de la Energía

Las conformaciones espaciales o puntos sobre la SEP más relevantes son aquellos que poseen la menor energía dado que corresponden a las conformaciones estables del sistema, es decir aquellas en las que es más probable encontrarlo y por ende son más representativas. Sin embargo, para un sistema de 3N coordenadas, existen numerosos mínimos locales sobre la superficie de energía potencial, lo que hace muy difícil identificar de manera exacta la geometría del sistema correspondiente a un mínimo para resolver este problema. En general, estos algoritmos permiten explorar la SEP hasta encontrar un punto donde cualquier movimiento en cualquier dirección haga que la energía del sistema crezca, lo que por definición es un mínimo. Entre los algoritmos más utilizados, están el de pasos descendentes y el del gradiente conjugado, cada uno con sus ventajas y desventajas al punto que en la práctica, se suele utilizar una combinación de ambos, ya que resultan complementarios entre sí. Por ejemplo, **primero se suele usar el método de pasos descendentes ya que se ha visto que funciona mejor cuando la estructura inicial difiere mucho de la/s estructura/s en el mínimo de la SEP y luego, cuando los contactos desfavorables ya han sido eliminados, se utiliza el método de gradiente**

conjugado que es más eficiente para explorar las estructuras cerca del mínimo de energía.

2.1.3. Dinámica Molecular (DM)

El proceso de relajar la estructura inicial mediante algoritmos de minimización (como se ha detallado en la sección previa) genera configuraciones individuales de mínima energía del sistema. Si bien en muchos casos, esta información puede ser suficiente para predecir de manera aproximada las propiedades del sistema (por ejemplo, para moléculas pequeñas o un conjunto de estas en fase gaseosa), la misma no es suficiente para comprender y predecir las propiedades de líquidos, soluciones y sólidos, así como para investigar el comportamiento de macromoléculas que poseen una SEP con numerosos mínimos locales muy cercanos en energía y donde la dinámica juega un rol fundamental.

En este punto, los métodos de simulación computacional nos permiten obtener una serie de configuraciones representativas del sistema, de modo que las propiedades termodinámicas extraídas del mismo se correspondan de manera precisa con los valores reales. Una de las técnicas más utilizadas es la dinámica molecular clásica en la cual se calcula la progresión real del sistema en función del tiempo. Para ello, las posiciones atómicas del sistema son propagadas temporalmente integrando las ecuaciones de movimiento de Newton para lo cual, existen numerosos algoritmos para realizar está integración de las ecuaciones de movimiento siendo el algoritmo de Verlet[14] o el de ”leapfrog”[15] de los más utilizados.

En cada paso de la dinámica, la energía y la fuerza sobre cada átomo son calculadas. Luego, en base a esto, se calculan cuáles serán las posiciones y velocidades de cada átomo en un intervalo de tiempo posterior (que por lo general, para el caso de proteínas, este intervalo de tiempo denominado *time step* (**dt**) está en el orden de los femtosegundos). De esta manera, se genera una trayectoria que describe como cambia la conformación del sistema en función del tiempo. Es importante notar que el dt debe ser tan pequeño como para poder describir de manera adecuada el movimiento natural del sistema, como si la integración de las ecuaciones de Newton fuera exacta. Sin embargo, cuanto menor sea el mismo, deberá realizarse un mayor número de cálculos para simular un tiempo total dado. En general, **una buena solución de compromiso, es utilizar un dt que sea 1/10 del tiempo característico del modo más rápido del sistema** (en biomoléculas, los modos de mayor frecuencia corresponden a la vibración de enlaces, en particular la vibración C-H y comúnmente se utiliza ”*time step*” de 2 femtosegundos[3]).

2.1.4. Entorno de Simulación: Condiciones Periódicas de Contorno

Actualmente, las simulaciones de DM se realizan en medios solvatados con moléculas de agua explícitas y rodeadas de otras moléculas de proteína como si estuvieran en un ambiente pseudo-cristalino o en una solución altamente concentrada. Estas condiciones periódicas permiten realizar simulaciones utilizando un número relativamente pequeño de partículas, pero de manera tal que las mismas experimenten fuerzas como sí estuvieran en condiciones de bulk. Para ello, el sistema se replica en todas direcciones para dar un arreglo de cajas periódicas que replica la caja central, en donde los átomos que están fuera, son simplemente imágenes de los átomos simulados en la misma. Este tipo de replica del sistema, es conocido como condiciones periódicas de contorno y asegura que todos los átomos simulados estén rodeados de átomos vecinos, así sea que sus vecinos sean imágenes replicadas en la caja aledaño o no.

2.1.5. Condiciones Experimentales: Temperatura y Presión

En las simulaciones por DM, es importante poder replicar las condiciones experimentales del sistema a simular, como por ejemplo, la presión y temperatura. En este punto, uno de los termostatos más utilizados en la simulación de sistemas grandes (de cientos o miles de átomos), como es el caso de las proteínas solvatadas con solventes explícitos, es el propuesto por Herman Berendsen en 1984[16] el cual, para mantener la temperatura y/o la presión constante en una simulación de dinámica molecular acopla débilmente el sistema a un baño externo a temperatura y/o presión fija con una constante de tiempo ajustable para el acoplamiento. Para ello, parte de la ecuación de movimiento de Langevin e impone la condición de acoplamiento global con la mínima perturbación localizada y modifica la misma dejando como remanente el acoplamiento global del sistema al baño. De esta manera, el termostato de Berendsen minimiza $\sum m_i(\Delta v_i)^2$ manteniendo $\sum(\Delta(\frac{1}{2}mv^2)$ constante con el objeto de que la forma Maxwelliana de la curva de distribución de velocidades se conserve (ver el trabajo de Berendsen et al. titulado ”Molecular dynamics with coupling to an external bath” para una demostración rigurosa de estás aproximaciones).

Es así como el baño resulta una fuente de energía térmica, pudiendo quitar o entregar calor al sistema según sea el caso y de este modo, las velocidades son reescaladas en cada paso de dinámica molecular, dado que la velocidad de cambio de temperatura es proporcional a la diferencia de temperatura entre el baño termostático y el sistema:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{\text{baño}} - T(t)) \tag{2.2}$$

en donde τ es un parámetro cuya magnitud determina el grado de acoplamiento entre baño

termostático y el sistema. Los cambios de temperatura entre los sucesivos intervalos de tiempo viene dada por:

$$\Delta T = \frac{\delta t}{\tau}(T_{\text{baño}} - T(t)) \tag{2.3}$$

Por consiguiente, el factor de escalamiento para las velocidades es:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}(\frac{T_{\text{baño}}}{T(t)} - 1) \tag{2.4}$$

En donde τ, es la constante de acoplamiento temporal, la cual determina la escala de tiempo para la cual la temperatura deseada es alcanzada, de esta manera cuando τ es grande el acoplamiento será débil y el sistema tenderá a comportarse de manera aislada, mientras que si τ es pequeño el acoplamiento será fuerte. **Lo recomendable es que el acoplamiento no sea ni muy débil como para que el sistema este aislado ni muy fuerte como para no permitir que la temperatura del sistema fluctué en torno a la temperatura deseada (para esto es conveniente mantener la relación $\frac{\delta t}{\tau}$ entre 0.001 y 0.002).**[3,6,16]

Por otro lado, el barostato de Berendsen se define de manera análoga según la siguiente ecuación:

$$\mu = [1 - \frac{\Delta t}{\tau_p}(P_0 - P)]^{\frac{1}{3}} \tag{2.5}$$

Donde τ_p es la constante de acoplamiento, P_0 es la presión de referencia, P la presión instantánea del sistema y Δt es el intervalo de tiempo de integración.

2.1.6. Modelos de agua explícita

Existen numerosos modelos de agua para simular el solvente que rodea a las biomoléculas de manera explícita. La mayor parte de los modelos de agua no polarizables asumen una carga positiva sobre los hidrógenos y una carga negativa junto con una interacción de Lennard-Jones en la posición del oxígeno. En general, se diferencian en la geometría del enlace (distancia HO-H y ángulo H-O-H), y la magnitud de la distribución de cargas (todos los modelos localizan la carga positiva sobre los átomos de hidrógeno pero difieren en la localización de la/s carga/s negativa/s y en las propiedades físicoquímicas del agua real que se pretende reproducir y contra las cuales se ajustan los parámetros del modelo).[17]

Entre los más populares se encuentran el modelo TIP3P (es el que fue utilizado en esta tesis), TIP4P, TIP5P, y TIP4P/2005. Todos estos modelos exhiben la misma geometría (la distancia O-H y el ángulo H-O-H en todos los modelos es el mismo que se encuentra en fase gaseosa, es decir 0.9572 Å y 104.53^0 respectivamente).

El modelo TIP3P fue propuesto por Jorgensen et al. en 1983 en el cual la carga negativa esta localizada sobre el átomo de oxígeno y las cargas positivas sobre los átomos de hidrógeno. Los parámetros del modelo (por ejemplo, aquellos correspondientes al

potencial de Lennard-Jones y las cargas sobre cada átomo) son ajustados de manera tal que reproduzcan el valor de entalpía de vaporización y la densidad del agua líquida en condiciones estándar de presión y temperatura y por lo general este modelo de aguas es el más ampliamente utilizado con campos de fuerzas parametrizados para biomoléculas.[18]

En el modelo de agua TIP4P, la carga negativa está localizada a 0.15 Å de la posición del oxígeno (usualmente llamado centro M), y si bien ha sido propuesto por Jorgensen et al., la geometría del modelo fue propuesta por Bernal y Fowler en 1933[19] y al igual que para el modelo de agua TIP3P, los parámetros del modelo de agua han sido ajustado de manera tal que reproduzcan la entalpía de vaporización y la densidad del líquido a temperatura ambiente. Si bien existen numerosos modelos de agua, un trabajo de Cerutti et al., en donde se compara la capacidad de los campos de fuerza de biomoléculas para reproducir las propiedades de las estructuras cristalográficas de proteínas en combinación con distintas tipos de modelos de aguas, llega a la conclusión de que **la elección del modelo de aguas tiene un efecto despreciable en comparación con la elección del tipo de campo de fuerzas utilizado para la proteína, y ya que estos han sido mayormente parametrizados utilizando un modelo de aguas de tipo TIP3P, es conveniente utilizar este modelo de aguas**.[20]

2.1.7. Validez del campo de fuerzas de AMBER en biomoléculas

Las predicciones cuantitativas de muchas propiedades de las proteínas a un nivel atomístico, dependen de manera crítica de la precisión del campo de fuerzas utilizado.[21] Esto requiere de una metodología rigurosa para poder poner a prueba los nuevos campos de fuerza, de manera tal que puedan reproducir propiedades biofísicas medidas experimentalmente, como los parámetros de relajación de spín por RMN, el acoplamiento escalar J, así como las estructuras cristalográficas.[22]

En el primer caso, se generan ensambles utilizando dinámica molecular basada en los distintos campos de fuerza clásicos (AMBER ff99, ff99SB, ff03) siendo los mismos evaluados a partir del cálculo de los corrimientos químicos utilizando los Cα, Cβ y C' en comparación con experimentos de RMN y en el segundo, se simulan con los distintos campos de fuerza y utilizando distintos modelos de agua la red cristalina de por ejemplo, la proteína toxina II (este método de prueba presenta la ventaja de poder identificar rápidamente estados conformacionales erróneos por simple comparación con la estructura cristalográfica).[20]
Con ambas metodologías de prueba, es posible observar que el campo de fueras ff99SB muestra la mejor perfomance para reproducir los ensambles observados experimentalmente (RMN) y es el que mejor mantiene las proporciones correctas de la celda unidad y reproduce la estructura y las fluctuaciones del *backbone* (como se observa en los experimentos de cristalografía a través del parámetro de fluctuación B-factor).

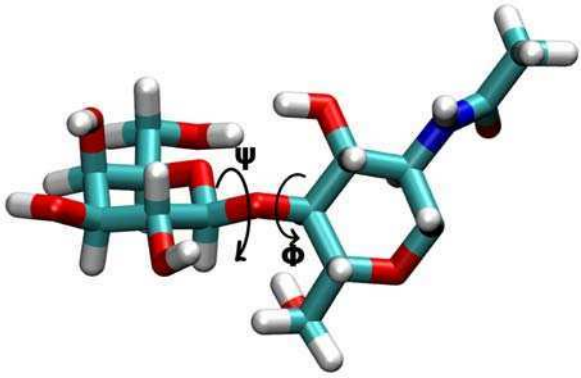

Figura 2.2. Representación esquemática de los ángulos de torsión entre monosacáridos vecinos

2.1.8. Campo de fuerza para Carbohidratos

El desarrollo de un campo de fuerzas confiable para poder simular el comportamiento de los carbohidratos en los distintos entornos representa actualmente un gran desafío, ya que es necesario considerar la influencia de la flexibilidad inherente de los glicanos en el desarrollo y validación de los parámetros. A diferencia de las proteínas y los oligonucleótidos (los cuales pueden ser reconstruidos como una combinación lineal de sus partes), los glicanos son frecuentemente estructuras ramificadas que en la mayoría de los casos, presentan carbonos quirales y que están unidos a un grupo hidroxilo, que pueden permanecer como tal o bien pueden formar una unión glicosídica para unirse a otro carbohidrato. Es decir, que si de la unión entre dos aminoácidos, solo dos dipéptidos pueden ser generados, de la unión de por ejemplo dos hexapiranosas, se pueden formar 20 disacáridos distintos.

Por otro lado, los carbohidratos presentan una peculiaridad adicional, dada por el efecto anomérico (que es una manifestación de la preferencia energética mostrada por los sustituyentes ricos en electrones unidos al carbono anomérico a posicionarse preferentemente en la posición axial respecto de la orientación ecuatorial), el cual tiene implicancias en azúcares de diferentes tamaños y en diferentes contextos.

En términos de la estructura de oligosacáridos, tiene importancia el efecto exoanomérico, que influye en la conformación del enlace glicosídico. Dado que este efecto tiene orígenes electrónicos y no puede ser descripto por un simple término de torsión, requiere una consideración extra para ser incluido correctamente en los campos de fuerza propios de carbohidratos. En relación con este enlace y de manera análoga a lo que sucede en proteínas, la conformación de los glicanos es generalmente definida por los ángulos de torsión entre los distintos monosacáridos vecinos, ϕ y ψ (ver figura 2.2). En general, los campos de fuerzas que tratan carbohidratos pueden ser clasificados de acuerdo a la filosofía con la que fueron diseñados. La mayor parte de ellos (CHARMM, GLYCAM06, GROMOS-

45A4, OPLS-AA-SEI, etc.) fueron desarrollados de manera tal que sean consistentes con las simulaciones de biomoléculas de gran tamaño, como por ejemplo, proteínas. El resto pueden ser identificados como campos de fuerzas para moléculas orgánicas (TRIPOS, MM3, MM4, etc) diseñados en general para representar moléculas orgánicas.

Todos estos parámetros computacionales requieren de un conjunto de datos experimentales rigurosos para poder validarlos y como ya ha sido mencionado, se suele utilizar los datos provistos a partir de experimentos de RMN. Sin embargo, existe una simbiosis entre los datos generados por RMN y el desarrollo de los campos de fuerza dado que a pesar de que los parámetros del campo de fuerzas deben ser validados frente a los resultados experimentales, es necesario tener modelos que permitan generar estructuras 3D capaces de explicar los resultados de RMN.[8,23] De los numerosos campos de fuerzas para carbohidratos existentes, cada uno con sus ventajas y desventajas, GLYCAM06 es uno de los más utilizados dado que es compatible con el campo de fuerzas de AMBER (ff99SB) presentando un amplio y completo conjunto de carbohidratos y glicoconjugados parametrizados[24] y que además ha sido utilizado en combinación con el campo de fuerzas de AMBER para estudiar un diverso rango de interacciones con proteínas, incluyendo aquellas relevantes a procesos infecciosos,[25] crecimiento celular[26] y conversión de biomasa.[27]

2.2. Sitios de hidratación (WS)

El cambio en la hidratación de dos moléculas que se unen ocurre en una escala molecular, con solamente un pequeño número de moléculas de agua que se afectan directamente. Sin embargo, las contribuciones de las mismas al proceso tiene un gran impacto, ya que cambios sutiles en la red de puentes de hidrógeno, están frecuentemente asociados con grandes cambios en la energía de interacción, con una ganancia de alrededor $10k_B$T por puente de hidrógeno (donde $k_BT \approx 0{,}6\ kcalmol^{-1}$). Como resultado de esto, las contribuciones de las moléculas de agua al proceso de unión de ligandos no es fácilmente cuantificable.[28] Adicionalmente, las moléculas de agua que solvatan regiones puramente hidrofóbicas, se comportarán de manera distinta y por lo tanto tendrán un efecto diferencial respecto de aquellas moléculas de agua que solvatan regiones puramente hidrofílicas.

De esta manera, existe un interés creciente en los últimos años por vincular la importancia de las moléculas de agua en el proceso de unión de ligandos y por desarrollar descriptores cuantitativos para los efectos de hidratación asociados al proceso de unión. Es interesante en este punto resaltar el trabajo de Baron, Setny y McCammon donde muestran, a partir de un modelo simple receptor-ligando utilizando simulaciones de dinámica molecular de una semi-esfera perfectamente hidrofóbica, solvatada con aguas explicitas a la que se une un ligando pequeño como metano, que la fuerza impulsora del procesos de unión viene

dada por la energía libre de desolvatar el ligando y el receptor (es decir que las interacciones directas entre el ligando y el bolsillo de unión no dominan el proceso de unión), y tal vez lo mas controversial es que el término de desolvatación viene dado principalmente por la componente entálpica, ya que se observa que las aguas en el seno hidrofóbico, tienen una entropía mayor que en el seno del solvente dadas las fluctuaciones de secado/mojado de la superficie hidrofóbica.[29,30] Por otro lado, solvatar superficies hidrofílicas pareciera tener el efecto opuesto, ya que como se ha visto, si bien las moléculas de agua están favorecidas desde el punto de vista entálpico a permanecer en la misma, desde el punto de visto entrópico se ven desfavorecidas, ya que pierden movilidad.[31–33]

Si pensamos ahora en la misma como una región heterogénea de zonas hidrofóbicas e hidrofílicas (en mayor o menor proporción dependiendo de la proteína), esto complejiza el análisis global de la estructura de solvatación en la superficie proteica. En el caso particular de las lectinas, si nos enfocamos en comprender la relación existente entre los grupos funcionales de los azúcares y la capacidad del solvente acuoso de mimetizar dicha interacción en el receptor libre, es relevante poner atención en las regiones hidrofílicas de la superficie proteíca que interactúan con el agua, ya que presentan residuos con grupos funcionales capaces de actuar como dadores o aceptores de puente de hidrógeno, ya sea con las moléculas de agua circundantes o bien con un potencial ligando que presente grupos polares en la orientación apropiada.

La interacción de una proteína con el solvente se puede estudiar a partir de la formación de los denominados sitios de hidratación (o **WS** por sus siglas en inglés ”Water Sites”) que **definimos como aquellas regiones del espacio con la capacidad de alojar moléculas de agua con mayor probabilidad que el seno del solvente**. Para identificar y caracterizar los WS y poder construir la estructura de solvatación de la superficie proteíca (o bien una región de interés, por ejemplo, el sitio de reconocimiento de carbohidratos (**CBS**) en lectinas), es necesario contar con una metodología apropiada para tal fin, la cual se describe en las secciones subsiguientes de este capítulo.

2.2.1. Identificación de los Sitios de hidratación (WS)

Con el objetivo de identificar la presencia de los WS en la superficie de la proteína, utilizamos la siguiente metodología.

En primer lugar, se seleccionaron los aminoácidos presentes en el CBS que presenten uno o más átomos capaces de actuar como dadores o aceptores de puentes de hidrógeno con el agua. Luego, se cálculo la función de distribución radial traslacional $g^{tr}(r)$ para los mismos, contando la presencia de moléculas de agua en un rango dr (alrededor de 0.1 Å) en torno a una distancia r de cada átomo. Cada una de estas funciones $g^{tr}(r)$ fue normalizada teniendo en cuenta el volumen accesible para las moléculas de agua el cual

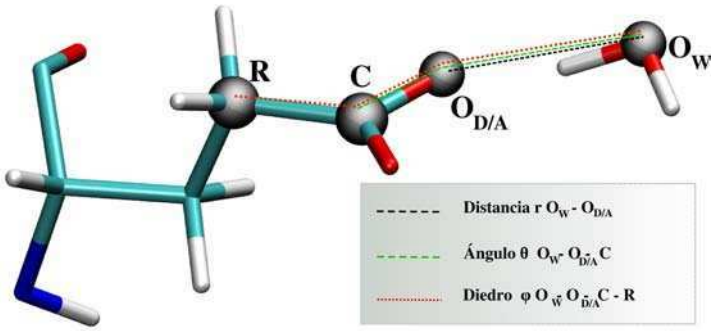

Figura 2.3. Definición del ángulo θ y el diedro ϕ. O_w representa el átomo de oxígeno de la molécula de agua y D/A, el átomo donor/aceptor de puente de hidrógeno respectivamente

fue calculado a partir de las funciones de distribución angular $g^{tr}(\theta)$ y $g^{tr}(\psi)$ teniendo en cuenta aquellas moléculas de agua, que se encuentran en la primera esfera de solvatación (es decir, aquellas que están incluidas en el primer pico del gráfico de $g^{tr}(r)$ -ver figura 2.4 A-) con respecto al átomo de referencia. Para ello, se contaron las moléculas de agua dentro de un intervalo dθ alrededor del ángulo θ y dϕ alrededor del ángulo ϕ, los cuales se definen como sigue: θ es el ángulo formado por los átomos de oxígeno de las moléculas de agua, el átomo donor/aceptor de puente de hidrógeno del aminoácido en cuestión y un átomo de referencia y ϕ es el dihedro definido por los átomos de oxígeno de las moléculas de agua, el átomo donor/aceptor del aminoácido y dos átomos de referencia, por lo general de los dos átomos de referencia uno está unido de manera directa al átomo donor/aceptor de puente de hidrógeno del aminoacido, y el otro al primero (ver figura 2.3 para más detalles) . Un análisis visual de los gráficos de distribución bidimensionales ψ vs. θ, permite identificar claramente los sitios de hidratación (WS) como aquellas regiones del espacio en donde la probabilidad de encontrar moléculas de agua es alta, mostrando como es de esperar que la distribución de moléculas de agua no es homogénea en el espacio sino que presentan orientaciones definidas ya que, la formación de puentes de hidrógeno entre el agua y los residuos de la proteína se establece con ángulos y distancias característicos.

A modo de ejemplo, la figura 2.4 A muestra, para el oxígeno carbonílico del Ácido Aspártico 16 (Asp16), perteneciente al CBS de Concanavalina A, el gráfico de distribución radial ($g^{tr}(r)$ vs. la distancia (**r**)) de los átomos de oxígeno de las moléculas de agua en torno a esa referencia y la figura 2.4 B , la distribución angular ψ vs. θ para los átomos de oxígeno de aquellas moléculas de agua que se encuentra en la primera capa de solvatación, es decir entre los 2.5 y los 3.2 Å aproximadamente desde el átomo de oxígeno carbonílico de Asp16. Como se desprende naturalmente a partir de la inspección visual de la figura 2.4 B, el gráfico angular bidimensional revela la presencia de dos regiones con alta probabilidad de encontrar moléculas de agua, es decir, dos WS bien definidos (identificados con flechas

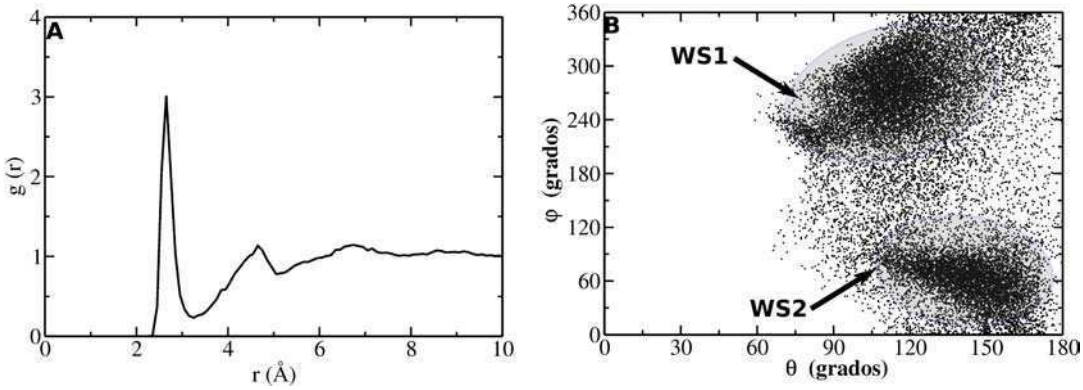

Figura 2.4. A) Función de distribución radial $g^{tr}(r)$ y B) gráfico bidimensional de distribución angular para los átomos de oxígeno de los moléculas de agua cercanas al grupo carboxilato de la cadena lateral de la Asp16 de Con-A

sobre la figura 2.4 B). Está metodología fue previamente aplicada en un trabajo previo del grupo para identificar los sitios de hidratación presentes en el CBS de Gal-1.[34] De esta manera, utilizando está estrategia de análisis para cada potencial WS, el centro de máxima probabilidad de encontrar moléculas de agua en la superficie proteica es determinado. Para los cálculos que siguen, estos puntos definen el centro del WS. Una molécula de agua se definirá como dentro de un dado WS, si la misma se encuentra a una distancia de 0.6 Å del centro del mismo (un valor que corresponde aproximadamente a un volumen arbitrario de 1 Å^3 para cada WS). Para eliminar ambigüedades en la definición de los WS, el análisis a partir de los gráficos bidimensionales fue complementado con un análisis tridimensional de los puntos de máxima probabilidad de los WSs utilizando un software de visualización y análisis de moléculas (**VMD** por sus siglas en inglés *V*isual Molecular Dynamics).[35]

Para chequear la convergencia de los parámetros reportados, también se computó lo mismo en diferentes segmentos de la Dinámica Molecular, o tomando diferentes referencias para definir un dado WS con el objeto de asegurar la reproducibilidad en la definición y en el cálculo de las propiedades de los mismos.

2.2.2. Cálculo de las propiedades termodinámicas y dinámicas de los sitios de hidratación (WS)

Con el objetivo de calcular la energía potencial asociada a la interacción de las moléculas de agua en los WS con la proteína y con el resto del solvente para cada sitio, las contribuciones a la energía electrostática y de vdW fueron calculadas entre las moléculas aguas que se consideran dentro de los WS con la proteína (E_p) y el resto del solvente (E_w), utilizando los parámetros y funciones del campo de fuerzas correspondientes. Las contribuciones fueron consideradas hasta una distancia de 8 Å desde el centro de los WS, dado que a este

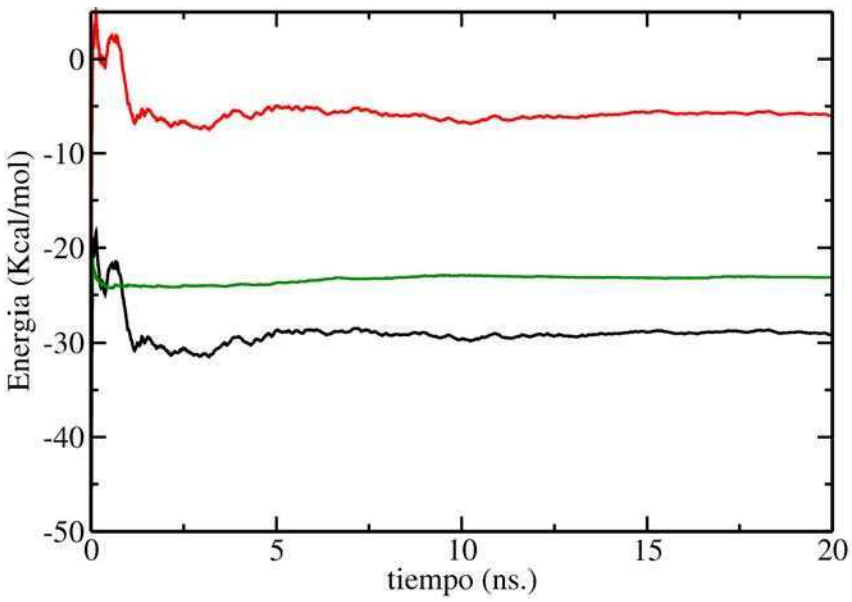

Figura 2.5. Energía de interacción promedio de las moléculas de agua dentro del WS con el resto del solvente (línea roja) , con la proteína (línea verde) y total (línea negra).

punto de corte la convergencia calculada es aceptable.[34] Para cada WS, se determinó la energía de interacción promedio total, denominada $< E_T >$ calculada como la suma de los promedios de $< E_p >$ y $< E_w >$. Por último, utilizando como referencia la energía de interacción promedio de una molécula de agua en el seno del solvente ($< E_{Agua-Agua} >$), la diferencia en la energía media de interacción de llevar una molécula de agua desde el seno del solvente hacia el WS, se calcula:

$$< \Delta E >=< E_T > - < E_{Agua-Agua} > \tag{2.6}$$

Está diferencia de energía corresponde al promedio de energía potencial que gana una molécula de agua cuando la misma es trasferida desde el seno del solvente hacia el WS correspondiente. Como puede observarse a modo de ejemplo en la figura 2.5, los valores de energía reportados para cada sitio de hidratación están razonablemente convergidos en la escala de tiempo que se utilizo para las simulaciones.

Para cada WS, se ha calculo también la probabilidad p(ν) de encontrar una molécula de agua dentro del mismo (en donde ν representa el tamaño del espacio utilizado para estimar la probabilidad), que luego fue normalizado con respecto a la probabilidad de encontrar una molécula de agua en el seno del solvente en el mismo volumen, lo que corresponde a la densidad del agua en la condiciones de presión y temperatura en la que fue llevada a cabo la simulación. De esta manera, lo que se informa es la probabilidad relativa de encontrar una molécula de agua dentro del WS respecto del seno del solvente, que llamaremos **WFP** (por

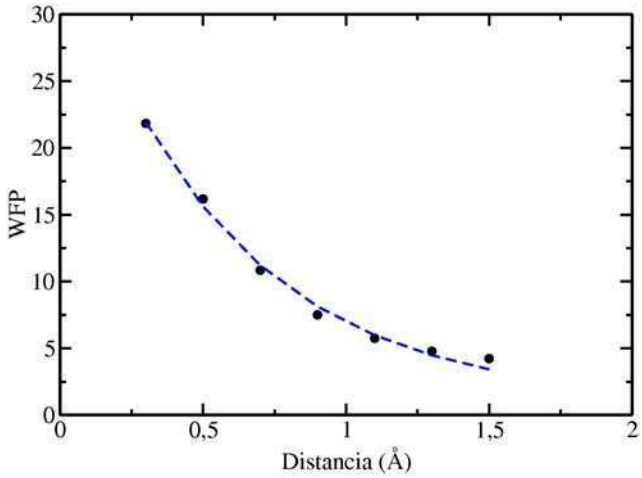

Figura 2.6. En puntos negros se observa la probabilidad relativa de aguas (WFP) vs. la distancia tomada desde el centro masas del sitio de hidratación (WS), perteneciente al WS9 al CBS de Concanavalina-A. La línea punteada azul representa el ajuste de una función mono exponencial

sus siglas en inglés *W*ater Finding Probability). Por ej., **si el WFP es 5, se debe interpretar como que es 5 veces más probable encontrar una molécula de agua en el espacio acotado por el WS respecto al seno del solvente**. En la figura 2.6 se muestra a modo de ejemplo, en puntos negros como cambia el valor de WFP vs. la distancia tomada desde el centro de masa del WS9 perteneciente al CBS de Concanavalina-A mientras en líneas punteadas azules se observa el resultado de un ajuste mono exponencial al comportamiento de WFP en función de la distancia. De esta manera, el valor de WFP que se informa es relativo al de un volumen arbitario de 1 $Å^3$ (dado que cada WSs se supone esférico, se toma, según la figura 2.6 el valor de WFP a 0.6 Å aproximadamente). Por otro lado, como una medida de cuán dispersas están las moléculas de agua que visitan un dado WS durante el tiempo de simulación, se ha computado el valor denominado R_{90}, que **representa la distancia medida desde el centro del WS, tal que el 90 % de las moléculas de agua que visitan el mismo se encuentren contenidas dentro de ese radio**. Por consiguiente, el valor de $\mathbf{R}_{90}$ calculado otorga una medida del tamaño del WS en unidades de volumen para cada sitio. El volumen de los WS, es también una medida indirecta de su dispersión y de ese modo, se lo puede asociar con la entropía traslacional, como se muestra en los trabajos de Li et al.[32,33]

Finalmente, para analizar la correlación entre la posición de los WS en el CBS en la proteína libre y la estructura del complejo proteína-carbohidrato, se calculó el parámetro R_{min} como **la distancia entre el centro de masa del WS y el átomo pesado más cercano del ligando en el complejo proteína-carbohidrato**.

Debido a que los WSs fueron calculados sobre la DM de la proteína libre y el cálculo de $\mathbf{R}_{min}$ es sobre la estructura del complejo, es necesario eliminar los grados de libertad

traslacionales que la proteína libre adquiere en la DM, y para ello se lo superpone sobre el complejo (que es tomado como estructura de referencia). Para ser consistentes en todos los casos, se tomo el criterio de que los aminoácidos pertenecientes al CBS (aquellos que se encuentran a menos de 5 Å del ligando en la estructura del complejo) se superpongan de manera tal que la desviación cuadrática media entre las coordenadas de los átomos (**RMSD** por sus siglas en inglés *Root Mean Square Deviation*) entre las distintas estructuras sea mínimo. Este procedimiento fue llevado a cabo con un modulo del programa *Visual Molecular Dynamics*.[35]

2.3. Método de clusterización para automatizar el análisis estructural y termodinámico de los WS

La principal desventaja del uso de las función de correlación, $g^{tr}(r)$ y los gráficos angulares para la identificación de los WS, es que el mismo es muy tedioso y artesanal, teniendo que interpretarse un número considerable de gráficos de distribución espacial, eliminando aquellos WS redundantes (es decir, que aparecen en más de un gráfico), plantear luego un posible esquema de la estructura de solvatación y finalmente con toda la información junta, volver a analizarla para evitar errores. Si bien la estrategia de análisis es robusta y se han desarrollado los programas que permiten llevarla adelante, la misma dificulta su extensión a muchos sistemas y es poco transferible a otros miembros de la comunidad científica que deseen utilizar este tipo de análisis para caracterizar la estructura del solvente concerniente a su proteína de interés. Esto ha motivado el desarrollo de un programa que permita alcanzar los resultados provistos mediante un método de análisis semi-automático. Para ello, decidimos utilizar una estrategia de clusterización de la posición espacial de las moléculas de agua a lo largo de la DM, tal como se describe a continuación.

2.3.1. Sitios de hidratación (WS) a partir de la estrategia de clusterización

Para encontrar los WS, se agrupan todas las moléculas de agua que interactúan con la superficie proteica en el lapso de la DM, de manera tal que el análisis combina un conjunto de conformeros de la proteína libre tomando ventanas de entre 1 y 2 ns. (tal que la estructura de la proteína estudiar no presente cambios conformacionales de mediana escala, como por ejemplo movimientos significativas de loops). Luego se toman las fotos del sistema espaciadas, entre 4 y 10 ps. y en una segunda instancia, es necesario relativizar el movimiento traslacional de la proteína alineando todos los residuos sobre los que se quiere hacer el análisis con respecto a una estructura de referencia (preferencialmente se alinea

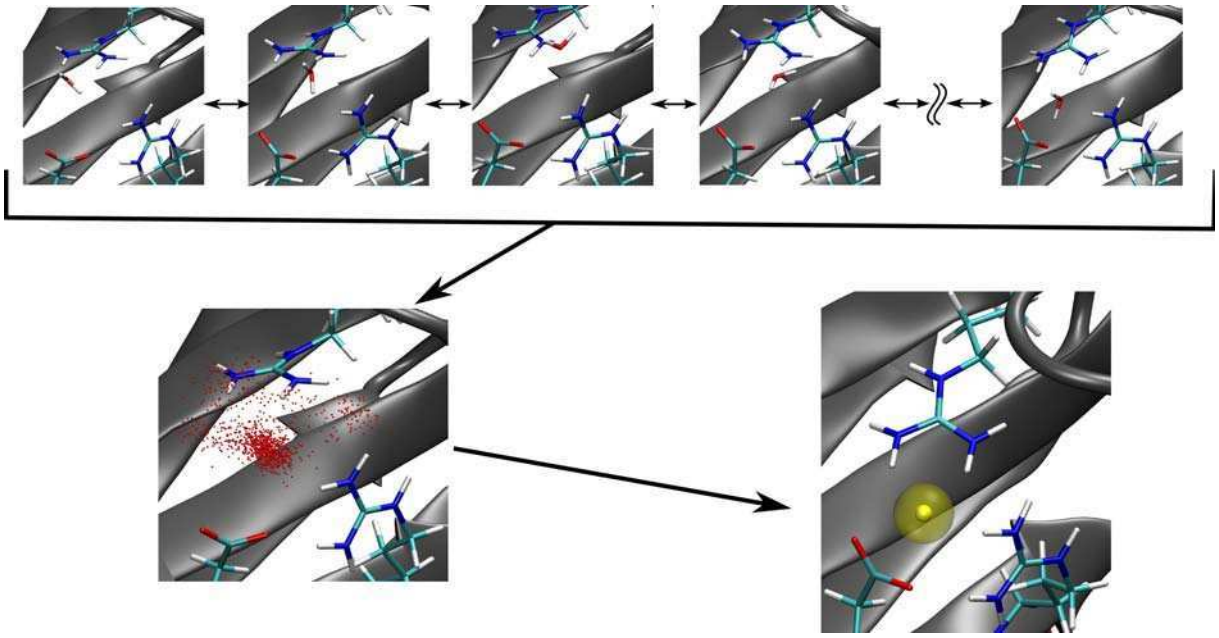

Figura 2.7. Esquema representativo de la determinación de sitios de hidratación (WS) en base al método de clusterización

respecto al *backbone* de la proteína con el objeto de maximizar el efecto de la movilidad de las cadenas laterales sobre la clusterización y para ello se utiliza un algoritmo de ajuste molecular, el cual requiere medir una diferencia numérica entre dos estructuras que están en diferentes posiciones en el espacio con el objetivo encontrar la orientación relativa de las moléculas en las que esta diferencia es minimizada.[36] De modo tal, que si se superponen todas las fotos de la dinámica molecular, se observan las regiones del espacio aledañas a la superficie proteica en donde la probabilidad de encontrar moléculas de aguas es máxima (ver panel inferior izquierdo de la figura 2.7), en donde se muestran los átomos de oxígeno de las moléculas de agua a menos de 4 Å de los residuos en cuestión (ya que como se ha observado en la bibliografía, así cómo en trabajos previos del grupo la primera capa de solvatación en las superficies proteicas no va más allá de 4 Å). Luego se aplica el algoritmo de clusterización, el cuál mide la distancia entre los átomos de oxígeno para las moléculas de agua en el CBS en las distintas fotos de la DM, detectando las regiones de alta ocupancia (para poder reconocer el cluster de aguas que se observa panel inferior izquierdo de la figura 2.7, es necesario ajustar los siguientes parámetros: **i)** distancia entre 2 átomos de oxígeno en fotos sucesivas, pudiendo discriminar aquellas que forman parte del WS de las que no (en general, este valor se estima entre 0.2 y 0.3 Å y **ii)** el número mínimo de aguas necesario para formar parte del cluster, que debe ser estimado en relación al número de moléculas de agua presentes en el seno del solvente a partir del valor de la densidad del agua a la temperatura de la simulación de dinámica molecular en un volumen de 1 $Å^3$ y se

lo relativiza al número de fotos). Este proceso se repite de manera cíclica hasta que tener en cuenta todas las moléculas de agua que visitan el CBS en el lapso de la DM.

2.3.2. Caracterización termodinámica de los sitios de hidratación (WS)

Una vez que se ha descripto la estructura de solvatación de la proteína como una suma de sitios de hidratación, es posible caracterizar termodinámicamente cada uno de ellos. En primer lugar, calcular la probabilidad relativa de encontrar aguas en el WS respecto del seno del solvente (WFP) para un volumen de 1 $Å^3$, resulta directo a partir de la definición de los WS descripto en la sección anterior, ya que una vez definido el WS, es un cálculo muy simple que consiste en contar las moléculas de agua que visitan dicho sitio en el transcurso de la DM y que cumplen con estar dentro de una esfera de volumen arbitrario de 1 $Å^3$, es decir a una distancia menor a 0.6 Å desde centro de de masa del WS. Además, como ya fuera definido, el R_{90} se determina como el radio que hay que tomar para incluir al 90 % de las moléculas de agua asignadas al WS dentro de una esfera definida por un radio con ese valor.

Por otro lado, una aproximación que se utiliza para poder calcular de manera más rigurosa la energía y la entropía de solvatación, es la propuesta por Lazaridis[37,38] en la conocida Teoría del Fluido Inhomogéneo (IFST), en la cual se trata al soluto como si estuviera fijo en el origen o bien, con respecto a una referencia generando un campo externo sobre el solvente en las cercanías de la superficie de interacción. De este modo, si se tiene en cuenta la dinámica del solvente tomando como referencia al soluto, se observarán regiones de distinta densidad a su alrededor, que podemos aislar en regiones acotadas del espacio tal que no alberguen a más de una molécula de agua por foto de la dinámica molecular (ver figura 2.8), lo que nos permite calcular la función de partición total de las moléculas de agua en esa región (solo las componentes rotacionales y traslacionales de las mismas, ya que el modelo de aguas que se suele utilizar es rígido y no contempla vibraciones entre los enlaces H-OH). Este formalismo, así como está planteado tiene la ventaja de aislar el efecto del soluto sobre las capas de solvente vecinos sin tener que tratar las grandes cantidades de solvente no perturbado que terminarían promediando hacia las propiedades del agua bulk, la delgada capa de solvente perturbado en la primera/segunda capa de solvatación de la superficie proteíca. Por consiguiente, la energía y la entropía de solvatación se pueden calcular como integrales de las funciones de correlación del solvente tomando al soluto como referencia.

En el marco de esta aproximación, la energía libre de solvatación puede ser descompuesta en 4 términos: La energía de interacción soluto-solvente (E_{SW}), la entropía soluto-solvente (S_{SW}), la energía de reorganización del solvente (ΔE_{WW}) y la entropía de reorganización del solvente (ΔS_{WW}). Estos dos últimos términos, reflejan la diferencia en las

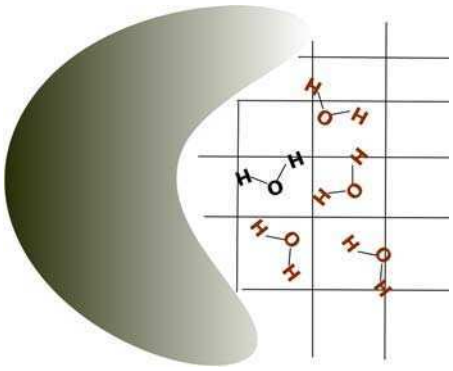

Figura 2.8. Representación esquematica de IFST

interacciones y correlaciones solvente-solvente de las aguas en el WS con respecto al seno del solvente (ver ecuación 2.7)

$$\Delta G_{solv} = E_{SW} + \Delta E_{WW} - T(S_{SW} + \Delta S_{WW}) \tag{2.7}$$

En donde, en la práctica, tanto el término E_{SW} como ΔE_{WW} es más fácil calcularlos de manera directa a partir de los resultados de la simulación, que en base a la integración de las funciones de correlación. El primer término, se obtiene a partir del promedio de la energía interacción de las moléculas de agua que visitan cada WS con el entorno proteico (se toman todos los átomos de la proteína a menos de 8 Å del agua en cuestión), mientras que el segundo término, se obtiene de la suma de la energía de interacción de la molécula de agua dentro del WS y el resto del solvente que la rodea, más el costo energético de extraer una molécula de agua desde el seno del solvente.

Por otro lado, dada la complejidad asociada a la naturaleza de las contribuciones entrópicas del solvente, este término no resulta trivial de calcular. No obstante, una posible solución que se ha planteado consiste en expandir la contribución entrópica de un fluido en una serie infinita de funciones de correlación multipartícula, con la mayor parte de las contribuciones provenientes de los términos de dos partículas (el cual puede ser fácilmente obtenido a partir de simulaciones de dinámica molecular), en donde se desprecian los términos superiores de tres a más partículas. Por lo tanto, de acuerdo a IFST, para cada WS S_{SW} se puede calcular por integración numérica de las funciones de correlación orientacionales y espaciales. De esta manera, las contribuciones entrópicas de primer orden para cada WS fueron calculadas según siguiente ecuación

$$S_{SW} = -k_b\rho_w/\Omega \int g_{sw}(\mathbf{r},\omega) ln(g_{sw}(\mathbf{r},\omega)) d\mathbf{r} d\omega \approx$$
$$-k_b\rho_w \int g_{sw}(\mathbf{r}) ln(g_{sw}(\mathbf{r})) dr - k_b N_W^V/\Omega \int g_{sw}(\omega) ln(g_{sw}(\omega)) d\omega$$

en donde **r** y ω son las coordenadas cartesianas de la posición y los ángulos de euler de la orientación de las moléculas de agua respectivamente; g_{sw} (**r**, ω) es la distribución roto-traslacional de las moléculas de agua con respecto al sistema de referencia fijo de la proteína; ρ_w es la densidad del agua pura; k_b es la constante de Boltzmann, Ω es el espacio orientacional accesible total para una molécula de agua y N_W^V es el número total de oxígenos de las aguas encontradas en un determinado WS en el volumen V, que se calcula a partir de la ecuación 2.8

$$N_W^V = \rho_w \int g_{sw}^{tr}(\mathbf{r}) d\mathbf{r} \tag{2.8}$$

De este modo, las funcion S_{SW} se integra numéricamente utilizando para la posición, coordenadas esféricas con un intervalo de 0.03 Å para **r** , 15 °para Θ y 30 °para ϕ y para cada uno de los tres ángulos de euler que describen la distribución rotacional, se hizo utilizando un intervalo de 10 °(estos últimos pueden ser calculados según el teorema de orientacional de Euler, el cual plantea que cada orientación en el espacio de un plano con respecto a otro puede ser descripta utilizando tres ángulos).

Finalmente, el término ΔS_{WW}, representa la entropía de reorganización del solvente y describe el cambio en la correlación solvente-solvente dada por la adición del soluto y simplemente representa el costo entrópico de remover una molécula de agua desde el seno del solvente (que según el trabajo de Lazaridis y Karplus,[39] $\Delta S_{WW} \approx 15{,}2 cal/molK$)

2.4. Docking

En general, las estrategias de docking tienen como objetivo predecir la estructura del complejo proteína-ligando en base a la estructura del receptor y del ligando separados. Las técnicas comúnmente utilizadas para realizar docking automático se pueden agrupar principalmente en dos categorías: métodos de coincidencia geométrica y métodos de simulación de formación de complejos o de docking. Los primeros crean un modelo del sitio activo, donde típicamente se incluyen zonas de interacción puente de hidrógeno y zonas que son estéricamente accesibles y por lo tanto tratan de ubicar un determinado inhibidor en una determinada conformación dentro del modelo, considerándolo usualmente como un cuerpo rígido (por ejemplo DOCK, el cuál es un programa muy eficiente para hacer un barrido rápido bibliotecas de compuestos).

La segunda clase de métodos de docking corresponden a aquellos que describen con mayor grado de detalle a la estructura del ligando y del receptor (por ej. el ligando se ubica de manera aleatoria fuera de la proteína, y luego explora traslaciones, orientaciones y conformaciones hasta que el estado ideal del complejo proteína-ligando es encontrado). Esta técnica es usualmente más lenta que los métodos de coincidencia geométrica, pero permiten que la flexibilidad del ligando sea considerada y pueda ser utilizada con mayor

grado de detalle para calcular la energía de unión del ligando en un sitio de unión plausible. En este punto, AUTODOCK es uno de los programas más ampliamente utilizados que combina una rápida manera de evaluar la energía de interacción proteína-ligando, en base a una grilla de puntos que tiene precalculadas las energías de interacción de a pares y que es usualmente utilizada durante toda la simulación con un algoritmo de búsqueda genético de tipo Lamarckiano.[40]

2.4.1. Introducción general a los esquemas de docking utilizados

Los algoritmos genéticos utilizan las ideas basadas en la genética y la biología evolutiva para explorar el espacio conformacional de ligando en el complejo. En el caso del docking molecular, el estado particular de una molécula se describe por un conjunto de valores que contienen información sobre la posición, la orientación y la conformación de un ligando con respecto a la proteína. Estos son los llamados estados variables del ligando y en los algoritmos genéticos cada estado variable es análogo a un gen. De este modo, el conjunto de estados del ligando conforman el genotipo, mientras que las coordenadas atómicas corresponden al fenotipo. En el docking molecular, el ajuste del ligando en una posición/conformación dada, "*fitness*", es la energía de interacción total del ligando con la proteína, y es evaluada utilizando la función de energía o puntaje específica calibrada empíricamente contra un conjunto de datos de ΔG_u medidos experimentalmente. De esta manera, la estrategia de búsqueda se puede plantear en base a un esquema híbrido de búsqueda global y luego local. Para abordar el primer punto se recurre a los algoritmos genéticos, los cuáles requieren generar una población inicial de ligandos con diferentes coordenadas (fenotipos), por lo general entre 150 a 300 individuos, dependiendo del tamaño del espacio de búsqueda, que van generando descendencia. Para ello, es relevante poder generar variabilidad entre las sucesivas generaciones y por lo tanto se imita un conocido proceso presente en la meiosis que es el llamado *crossing over* o intercambio cruzado, en el cuál para el caso del docking molecular, se combinan pares aleatorios de individuos de una población entre los cuales se intercambian genes para dar lugar a un nuevo individuo, es decir, un ligando con un dado estado posicional y conformacional, que conlleva una mezcla aleatoria de algunos genes parenterales. Además, algunos individuos de la descendencia son sometidos a mutaciones aleatorias en las cuales un gen cambia su estado por un valor aleatorio (es decir, que en la práctica, en los estados hijos el ligando presenta conformaciones y posiciones que difieren de los parenterales en algún aspecto). Por último, la selección de la descendencia de la siguiente generación se hace en base a la función de energía, es decir, que solo sobrevive aquel individuo que se encuentra en el entorno más favorable (en otras palabras, aquel que tiene el valor de energía de interacción proteína-ligando más negativo). Con respecto a la búsqueda local, se realiza una

minimización de energía utilizando el algoritmo de *Solis and Wet*, el cual tiene la ventaja de no requerir la información del gradiente de energía provisto a través de la superficie de energía potencial, lo que facilita la búsqueda en el espacio torsional.[41] Este método híbrido, que combina un algoritmo genético con un método adaptativo de búsqueda local es conocido como Algoritmo Genético Lamarkiano con búsqueda local y que a diferencia de otros métodos de exploración que han sido ampliamente utilizados como Monte Carlo o Simulated Annealing, permite abordar problemas en los se utilizan un número mayor de grados de libertad. Además, para hacer las minimizaciones, es posible utilizar el mismo campo de fuerzas que se utiliza para hacer el cálculo de la energía de interacción proteína-ligando.

Como ya se menciono, existen muchos métodos basados en los campos de fuerzas clásicos, que han mostrado excelentes resultados para estudiar procesos moleculares en función del tiempo, que funcionan bien para optimizar conformaciones, realizar cálculos de energía libre entre moléculas introduciendo cambios sutiles a nivel de átomos, pero tienen la desventaja de que son muy costosos computacionalmente, y desafortunadamente no muestran una gran eficiencia para ordenar correctamente la afinidad de compuestos por un dado receptor que difieren en un número significativo de átomos entre sí. Para compensar las deficiencias tradicionales de los campos de fuerza, se puede modelar la energía libre de unión proteína-ligando en base a una expansión de la conocida expresión de energía libre de Gibbs agregando términos que compensan o estiman los efectos entrópicos y del solvente, por ejemplo como se muestra en la ecuación 2.9

$$\Delta G = \Delta G_{VDW} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol} \tag{2.9}$$

en donde los 5 términos del campo de fuerza incluyen dispersión-repulsión, unión hidrógeno, electrostático y desviación de la geometría óptima en las uniones covalentes (términos de unión), respectivamente. Por su parte, ΔG_{tor} modela la restricción de las rotaciones internas, rotaciones globales y traslaciones. Por último, ΔG_{sol} modela la desolvatación luego de la unión del ligando y tiene en cuenta el efecto hidrofóbico (son los cambios entrópicos del solvente en la interfase proteína-ligando).

En particular, Autodock implementa una aproximación similar utilizando el ciclo termodinámico de Wesson y Eisenberg.[42] La función incluye 5 términos (ver ecuación 2.10)

$$\Delta G = \Delta G_{VDW} \sum_{i,j}^{n} \left(\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j}^{n} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} + \frac{D_{ij}}{r_{ij}^{10}} \right) + \Delta G_{elec} \sum_{i,j}^{n} \frac{q_i q_j}{\epsilon_{ij} r_{ij}} + \Delta G_{tor} N_{tor} +$$

$$\Delta G_{sol} \sum_{i,j}^{n} (S_i V_j + S_j V_i)\, e^{r_{ij}^2/2\sigma^2} \tag{2.10}$$

en donde los 5 términos de ΔG de la derecha (ΔG_{VDW}, ΔG_{hbond}, ΔG_{elec}, ΔG_{tor} y ΔG_{sol}) son coeficientes que se ajustan contra datos experimentales de un conjunto de complejos cuyas constantes de disociación son conocidas. Las sumatorias se llevan a cabo sobre todos los pares de átomos del ligando, i, y de la proteína, j, además de todos los pares de átomos del ligando que están separados por tres o más enlaces. Los tres primeros términos incluyen las contribuciones en vacío: Un término dispersivo 12-6 de Lennard-Jones, un término direccional 12-10 que incluye la interacción de puente hidrógeno, donde E(t) es un término direccional que pesa el ángulo, t, entre el átomo de prueba y el átomo blanco y por último, un término que tiene en cuenta el potencial electrostático (el cuál utiliza cargas atómicas parciales sobre cada átomo estimadas usualmente con el método de Kollman o Gastaiger[43,44]). El cuarto término es el que incluye la contribución a la energía libre dada por la desolvatación, el cual está basado en el método de Stouten et al.,[45] que presenta la ventaja de funcionar correctamente con un modelo de cálculo de energía basado en el precálculo de una grilla de afinidad, como lo hace Autodock. En este método, para cada átomo en el ligando, un volumen parcial de los átomos de la proteína que lo rodean son pesados por una función exponencial y luego sumados, evaluando el porcentaje del volumen alrededor de los átomos del ligando que es ocupado por los átomos de la proteína. Este porcentaje es luego pesado por los parámetros de solvatación de los átomos del ligando para dar la energía de desolvatación. Como una medida de la penalidad entrópica dada por la restricción roto-traslacional que se produce al unirse el ligando al receptor, se incluye un término que es proporcional a N_{tor} (que es el número de enlaces en el ligando que pueden ser descriptos por una hibridación de tipo sp^3) y que es constante durante todo el cálculo de docking ya que depende principalmente de los grados de libertad rotacionales del ligando que fueran previamente definidos por el usuario (por lo general no más de 12). Finalmente, utilizando un método de interpolación trilineal se evalúa rápidamente cada término energético para cada tipo de átomo del ligando en una dada posición, utilizando los mapas de energía que fueran previamente calculados sobre el entorno proteico.

Para la construcción de los mapas de energía, Autodock genera una grilla tridimensional en torno a las coordenadas de los residuos de la proteína. Un átomo de prueba visita cada punto de la grilla, se calcula la energía de interacción y la guarda. La grilla completa de interacciones es almacenada en una tabla y la energía se evalúa rápidamente. De esta manera, el mapa de energía de cada tipo de átomo del ligando se guarda en tablas separadas, incluyendo las energías de dispersión/repulsión y de formación de puentes de hidrógeno. Además se calcula una grilla de interacción electrostática, utilizando una partícula de prueba con una carga positiva y una grilla que incluye las energías de desolvatación

2.4.2. Protocolo y método de docking Sesgado (WSBDM)

Basándonos en la hipótesis que se plantea en el capítulo 3 de que los grupos -OH tienden a ocupar o reemplazar la posición de las moléculas de agua que son retenidas por los residuos del CBS, las cuales están representadas por la posición de los WSs, se ha modificado el esquema de docking convencional de Autodock, incluyendo esta información, en lo que hemos llamado método de docking sesgado o **WSBDM**, (por sus siglas en inglés *W*ater Sites Bias Docking Method).

Para ello, se ha modificado la función de energía de Autodock4, agregando un término adicional para cada átomo de oxígeno del ligando (en la nomenclatura de Autodock4 estos átomos son llamados OA) de manera que la función original que describe la interacción de los átomos de oxígeno queda como se describe en la ecuación 2.11,

$$\Delta G_O^M = \Delta G_O^{AD} - RT \sum_{i=1}^{N} ln(WFP_i) exp(\frac{-((x - x_{WS,i})^2 + (y - y_{WS,i})^2 + (z - z_{WS,i})^2)^{\frac{1}{2}}}{R_{90,i}}) \tag{2.11}$$

donde ΔG_O^M corresponde al valor de la función de energía modificada para los átomos de oxígeno del ligando, ΔG_O^{AD} corresponde al valor de energía original, WFP_i es el valor de probabilidad relativa de las moléculas de agua dentro del WS respecto del seno del solvente, como fuera previamente definido, del "i-ésimo" WS considerado. $x_{WS,i}$, $y_{WS,i}$ y $z_{WS,i}$ son las posiciones correspondientes de los WSs y x, y y z son las coordenadas de los puntos de la grilla. Por último, $R_{90,i}$ es el R_{90} correspondiente al WSs i-ésimo. De esta manera, cada WS considerado, provee un término de interacción energético entre el centro del WS y cada átomo de oxígeno del ligando (por ejemplo, cada oxígeno de un grupo hidroxilo del carbohidrato), con una magnitud que es directamente proporcional al ln(WFP) y cuya amplitud está relacionada con el tamaño del WS a través de R_{90}.

Está función, cómo se verá más adelante, está inspirada en los resultados presentados en los capítulos 3 y 4, en donde se mostrará que los parámetros de los WSs que mejor correlacionan con la posición que irán a ocupar los grupos -OH de los carbohidratos, medidos a partir del valor de R_{min}, son WFP y R_{90}.[46,47] Por otro lado, la forma funcional de la penalidad impuesta a la modificación de la energía original esta basada en que la probabilidad de encontrar una molécula de agua en un dado WS decae exponencialmente con la distancia al centro de masa del mismo (ver figura 2.6). A modo de ejemplo, en la figura 2.9 se muestran los mapas isosuperficie para distintos valores de energía construidos a partir del método convencional (izquierda) y con el método modificado, (derecha), para el CBS de Galectina 3. La figura muestra claramente que el mapa de energía para el método convencional **CADM** (por sus siglas en inglés "*C*onventional Autodock Docking Method") y el método sesgado es muy similar para un valor de -0.5 Kcal/mol. Sin embargo, a medida

que la energía se hace más negativa, el mapa de energía para el método sesgado **WSBDM** (por sus siglas en inglés "*W*ater sites Bias Docking Method") muestra pozos de energía en los lugares en donde la probabilidad de encontrar moléculas de agua es alta, WFP, los cuales no están presentes en la grilla original (ver esferas amarillas sobre el panel derecho de la figura 2.9). De esta manera, para los cálculos de docking, el WSBDM es utilizado de la misma manera que el método convencional pero modificando la grilla original con la información de la posición de los WSs previamente calculados. Para una estricta comparación de ambos métodos, todos los otros parámetros utilizados en los esquemas de docking (los cuales se detallan en la sección Metodología Especifica del capítulo 5), como el tamaño y posición de las grillas, el número de las corridas de docking, se mantuvieron iguales.

2.5. Energía Libre de unión

La predicción teórica de la energía libre de unión (ΔG_u), es uno de los desafíos más importantes en bioquímica computacional y existen actualmente numerosos métodos que permiten estimarla con diferentes grados de aproximación. Entre ellos, tenemos los conocidos métodos de punto final, que calculan el ΔG_u sobre la base del post-procesamiento de los resultados que se extraen de la simulación por DM y que recientemente han mostrado notables avances y mejoras en la estimación del ΔG_u, utilizando entre otras técnicas de análisis basadas en la mecánica cuántica y en la inclusión explicita del solvente. En segundo lugar, tenemos los métodos que están basados, no en los estados finales del proceso de unión, sino en el proceso en sí, como por ejemplo, aquellos métodos basados en dinámicas de no equilibrio sesgadas, de muestreo de asociación/disociación o en transformaciones de estado libre/unido como en el método de integración termodinámica.

Cada uno de estos métodos tiene sus ventajas y desventajas. Por ejemplo, los métodos de punto final son más rápidos que los de asociación/disociación e integración termodinámica, aunque son menos precisos y dependen mucho del sistema cuando se utilizan aproximaciones clásicas. Se puede alcanzar una mejora significativa cuando se los combina con métodos cuánticos o se incluyen efectos de solvente explícito (dada la posibilidad de combinar los resultados del análisis de la solvatación explícita y las metodologías basadas en cálculos cuánticos, los métodos de punto final ofrecen una clara y directa vía para mejorar su precisión, lo que puede ser fácilmente incorporado en un programa de diseño racional de drogas o *virtual screening*). Otra ventaja de los métodos de punto final es que los cálculos requeridos son relativamente fáciles de llevar a cabo y pueden ser automatizados y analizados, mientras que en los métodos de asociación/disociación o integración termodinámica, la definición de la coordenada de reacción debe ser casi siempre determinada y el tiempo de análisis de los datos crudos para obtener el ΔG_u suele demandar mucho

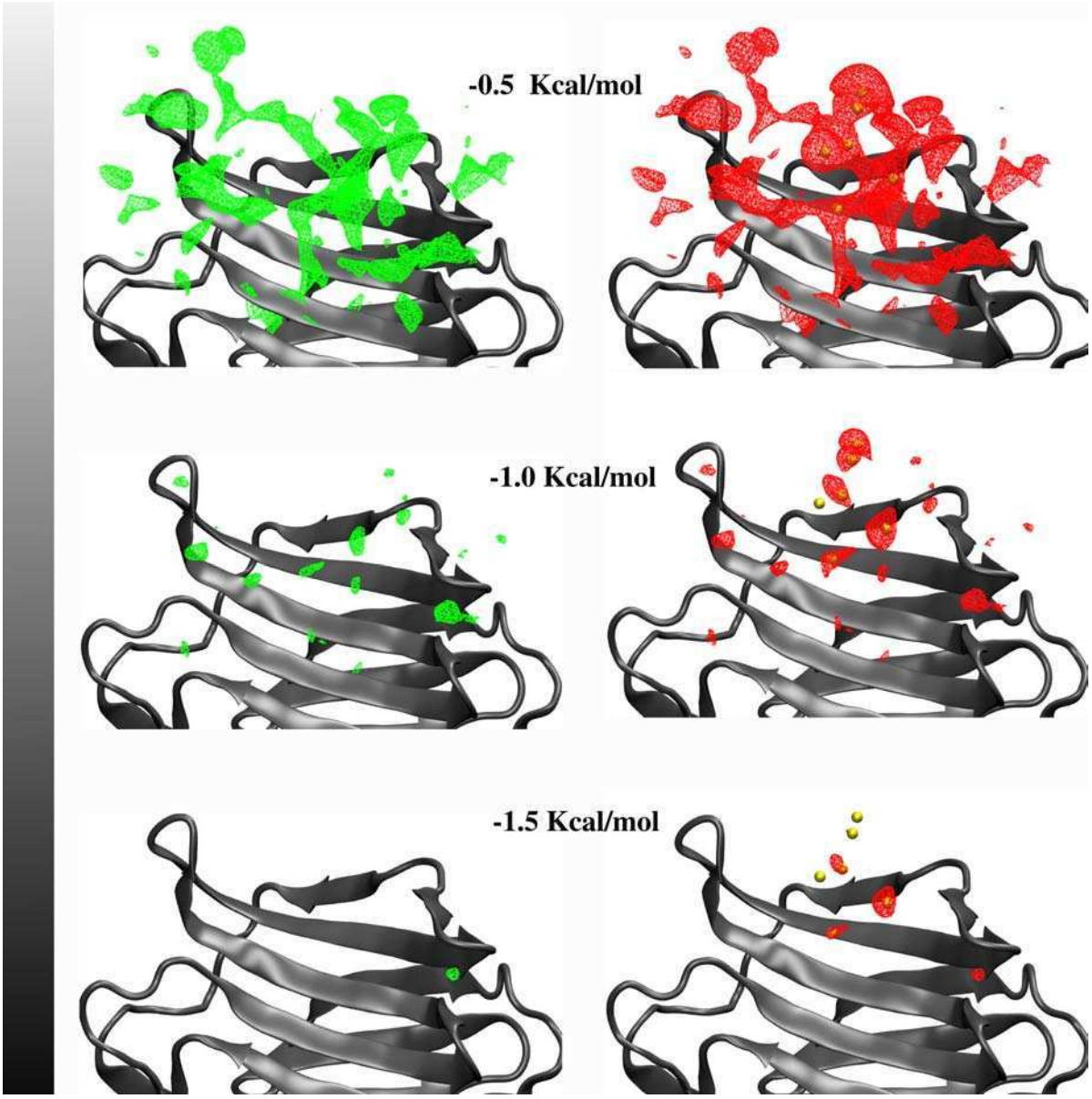

Figura 2.9. Representación de mapas de isoenergéticos para un átomo de prueba de tipo oxígeno, sobre una grilla construida sobre el CBS de Gal-3, para el método de docking convencional (CADM) -izquierda-, y modificado (WSBD) - derecha-.

tiempo. Por último, una de las desventajas de los métodos de punto final es que requieren del conocimiento de la estructura del complejo proteína-ligando, y por consiguiente, si no se cuenta con esta información, no puede realizarse la DM del complejo o bien, la estructura del mismo tiene que ser aproximada en base a un método de muestreo, como por ejemplo un método de docking.

2.5.1. Métodos de punto final

Los métodos de estimación de energía libre de punto final encierran un conjunto diverso de estrategias que están basadas en la simulación de únicamente el complejo proteína-ligando, y muchas veces la proteína y el ligando libre o no unido, para determinar el ΔG_u entre ellos. Usualmente, los métodos computacionales intentan estimar la afinidad de unión relativa entre un conjunto de compuestos relacionados estructuralmente, más que su valor absoluto. Este tipo de metodología generalmente involucra la realización de una simulación de dinámica molecular del complejo y/o del receptor y del ligando en solución y están basados principalmente en el post-procesamiento de estás trayectorias de dinámica molecular. Los métodos de punto final están basados en una determinación de la Energía Libre de Unión, como una suma de diferentes contribuciones energéticas y entrópicas (como se muestra en la ecuación 2.12).[6]

$$\Delta G_u = \Delta E_{int} - T\Delta S_{Lig} - T\Delta S_{prot} + \Delta G_{sv} \tag{2.12}$$

Donde ΔE_{int} corresponde directamente a la energía de interacción proteína-ligando en el complejo, ΔS_{Lig} da cuenta de la pérdida entrópica del ligando luego de la unión al receptor y ΔS_{prot} corresponde al cambio en la entropía conformacional de la proteína luego de la unión. ΔG_{SV} representa el cambio de energía libre del solvente luego de la unión y T es la temperatura del sistema. De estas contribuciones, usualmente las más importantes son ΔE_{int} y ΔG_{SV} y por consiguiente existen actualmente diversas estrategias para calcularlos de manera aproximada. Es importante notar que una aproximación inherente a esta clase de métodos, es que supone que todo el espacio conformacional del complejo proteína-ligando es explorado en la trayectoria de Dinámica Molecular, desde el cuál cada término es calculado.[48]

2.5.2. Métodos clásicos con solvente implícito: MM-PB(SA) y MM-GB(SA)

Posiblemente los métodos de punto final más utilizados y conocidos para calcular la energía libre de unión son MM-PB(SA) (*Molecular Mechanics Poisson-Boltzmann Surface*

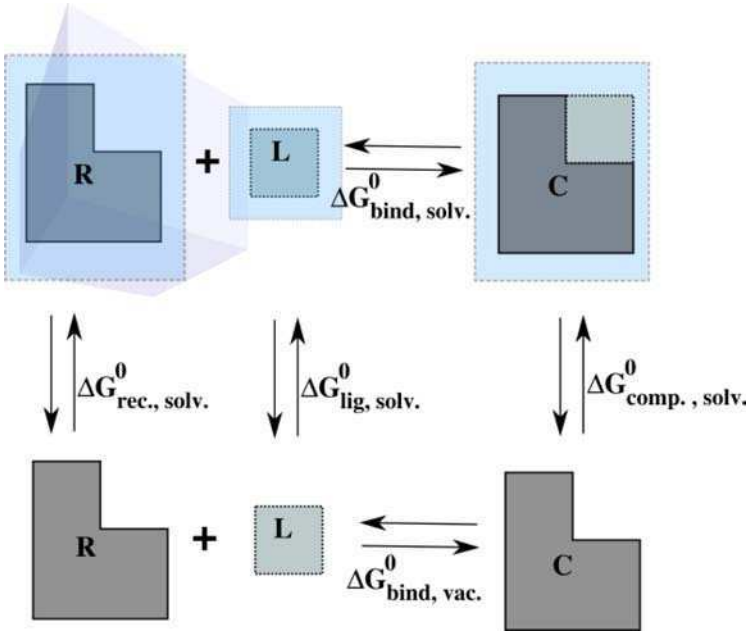

Figura 2.10. Ciclo termodinámico para el proceso de unión del ligando (L) sobre un dado receptor (R) dando como resultado el complejo (C). El ciclo permite separar los distintos eventos del proceso de unión y asociarlos con la energía libre de unión en vacío, ($\Delta G_{u,vacio}$), panel inferior, de la energía de solvatación en cada subsistema R, L y C ($\Delta G_{R/L/C,solv.}$)

Area) y MM-GB(SA) (*Molecular Mechanics-Generalized Born Surface Area*), que combinan un campo de fuerzas clásico con un modelo de solvente continuo para el cálculo de ΔE_{int} y ΔG_{SV} respectivamente. Para determinar la energía libre de unión con este tipo de métodos, el procedimiento es muy simple (como se muestra esquemáticamente en la figura 2.10). En primer lugar, es recomendable realizar una dinámica molecular larga para cada complejo proteína-ligando (en general entre 20 y 50 ns. dependiendo del sistema). Luego, se debe generar un conjunto de entre 1000 y 5000 fotos del sistema seleccionadas de manera aleatoria, o bien fotos consecutivas pero separadas como mínimo cada 4 ps.,[49] luego se remueve el solvente explícito, y finalmente a partir de la dinámica del complejo, se construyen tres trayectorias (la del complejo, la del receptor y la de ligando). Una vez que estos nuevos conjuntos de fotos son generados, se cálcula el ΔE_{int} de la ecuación 2.12 como el promedio de la diferencia en la energía potencial derivada de la evaluación de cada estructura del sistema, utilizando el correspondiente campo de fuerzas, como se muestra en la ecuación 2.13.

$$\Delta E_{int} = E_{complex} - E_{Receptor} - E_{Ligando} \tag{2.13}$$

Si se recurre a la aproximación de una trayectoria (es decir, solo se realiza la trayectoria del complejo y en base a esta se construyen las trayectorias para el receptor y el ligando), como se ha hecho en está tesis, en el cálculo de ΔE todos los términos de unión se cancelan entre

sí y el único aporte que se tiene a ΔE_{int} viene de los términos de no unión E_{vdW} y E_{elec}.

Una ventaja de la aproximación de trayectoria simple para los cálculos de energía libre de unión, es que la convergencia de los resultados se alcanza más rápido y se disminuye la dispersión de los mismos ya que las diferencias conformacionales que se generan en regiones lejanas al sitio de unión para las biomoléculas (en particular para el receptor y la porción proteica del complejo) no se cancelan cuando el cálculo se hace a partir de tres dinámicas moleculares generadas de manera independiente, una para el complejo y otras para el receptor y ligando, respectivamente.[48] Finalmente, las fotos derivadas de la aproximación de una trayectoria se utilizan para calcular la contribución del ΔG_{sv} a la energía de unión global, que puede ser separada en dos términos según la ecuación 2.14,

$$\Delta G^{solv} = \Delta G^{solv}_{pol} + \Delta G^{solv}_{np} \tag{2.14}$$

En donde el primero corresponde al componente polar y el segundo al no polar de ΔG^{solv}, respectivamente. Por un lado, la contribución polar al término de solvatación se calcula utilizando el modelo de PB o GB respectivamente, y el término no polar incluye un término de cavitación que tiene en cuenta el efecto hidrofóbico de la desolvatación cuando se forma el complejo a partir de sus componentes libres en solución. Por ejemplo, el cambio en el área expuesta a solvente cuando se forma el complejo (la cual es simplemente calculada como el producto del coeficiente de tensión superficial por el cambio de área expuesta a solvente ($\Delta SASA$) luego de la formación del complejo). Asumiendo que el solvente puede ser bien representado por un dieléctrico continuo, la ecuación de Poisson-Boltzmann (PB) es el método físicamente correcto para determinar el ΔG^{solv}_{pol} y que ha sido ampliamente utilizado.[50–52] La aplicación de PB a geometrías moleculares requiere resolver de manera numérica ecuaciones diferenciales de segundo orden, lo que es computacionalmente costoso. Alternativamente, el modelo generalizado de Born (GB) se ha vuelto muy popular en los últimos años como una aproximación computacionalmente más eficiente a las soluciones numéricas de la ecuación de PB, especialmente para su uso en dinámica molecular.[53–63]

En general, los modelos basados en la aproximación de GB, evalúan el término polar de la energía libre de solvatación como una suma de pares de interacción entre las cargas atómicas. Cuando la constante dieléctrica del soluto es 1 y la del solvente es mucho más grande (en particular, 80 para solventes acuosos), las interacciones pueden describirse de manera aproximada por la función analítica propuesta por Still et al., la cual hace una interpolación entre el límite coulómbico a distancias largas y el limite de Born u Onsager a distancias cortas[64] de acuerdo a la ecuación 2.15,

$$\Delta G_{pol} \approx \Delta G_{GB} = -1/2 \sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)}} \left(1 - \frac{1}{\epsilon_w}\right) \tag{2.15}$$

donde r_{ij} es la distancia entre los átomos i y j, q_i y q_j son las cargas parciales sobre cada átomo y ϵ_w es la constante dieléctrica del solvente. Los parámetros claves en esta expresión de la función de GB son el radio efectivo de interacción entre átomos, R_i y R_j, los cuales representan el grado de ocluímiento de cada átomo al solvente, y el cual es uno de los parámetros más influyentes en los solventes continuos. Es interesante en este punto mencionar el trabajo de Tingjun Hou et al., donde muestran una comparación sistemática para el cálculo de la energía libre de un conjunto de complejos proteína-ligando entre MM-PB(SA) y MM-GB(SA), **mostrando que si bien MM-PB(SA) funciona mejor para predecir el valor absoluto de la energía libre, MM-GB(SA) es capaz de predecir mejor la afinidad relativa entre complejos**. Considerando que MM-GB(SA) es computacionalmente menos costoso, este método surge como una promisoria herramienta para ser utilizada en el diseño de drogas, en donde el correcto orden por afinidad de las drogas propuestas es un factor determinante.[65]

Referencias

[1] David a Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The Amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–88, December 2005.

[2] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, 2002. cited By (since 1996)722.

[3] S A Adcock and J A McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5):1589–1615, 2006.

[4] L Nilsson, G M Clore, A M Gronenborn, A T Brünger, and M Karplus. Structure refinement of oligonucleotides by molecular dynamics with nuclear overhauser effect interproton distance restraints: Application to 5′ d(C-G-T-A-C-G)2. *Journal of Molecular Biology*, 188(3):455–475, 1986.

[5] J O Wrabl, D Shortle, and T B Woolf. Correlation between changes in nuclear magnetic resonance order parameters and conformational entropy: Molecular dynamics simulations of native and denatured staphylococcal nuclease. *Proteins: Structure, Function and Genetics*, 38(2):123–133, 2000.

[6] Andrew Leach. *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall, 2 edition, 2001.

[7] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, New York, NY, USA, 1989.

[8] E Fadda and R J Woods. Molecular simulations of carbohydrates and protein-carbohydrate interactions: Motivation, issues and prospects. *Drug Discovery Today*, 15(15-16):596–609, 2010.

[9] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983. cited By (since 1996)8477.

[10] W.R.P. Scott, P.H. Hünenberger, I.G. Tironi, A.E. Mark, S.R. Billeter, J. Fennen, A.E. Torda, T. Huber, P. Krüger, and W.F. Van Gunsteren. The gromos biomolecular simulation program package. *Journal of Physical Chemistry A*, 103(19):3596–3607, 1999. cited By (since 1996)705.

[11] P.K. Weiner and P.A. Kollman. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *J. Comput. Chem.*, 2:287–303, 1981. cited By (since 1996)348.

[12] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz Jr., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995. cited By (since 1996)7140.

[13] C.I. Bayly, P. Cieplak, W.D. Cornell, and P.A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The resp model. *Journal of Physical Chemistry*, 97(40):10269–10280, 1993. cited By (since 1996)2093.

[14] L. Verlet. Computer experiments on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159(1):98–103, 1967. cited By (since 1996)2595.

[15] R.W. Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, 9:136–211, 1970. cited By (since 1996)492.

[16] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, A. Dinola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984. cited By (since 1996)9908.

[17] C. Vega, J.L.F. Abascal, M.M. Conde, and J.L. Aragones. What ice can teach us about water interactions: A critical comparison of the performance of different water models. *Faraday Discussions*, 141:251–276, 2008. cited By (since 1996)93.

[18] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. cited By (since 1996)10978.

[19] J.D. Bernal and R.H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *The Journal of Chemical Physics*, 1(8):515–548, 1933. cited By (since 1996)943.

[20] D.S. Cerutti, P.L. Freddolino, R.E. Duke Jr., and D.A. Case. Simulations of a protein crystal with a high resolution x-ray structure: Evaluation of force fields and water models. *Journal of Physical Chemistry B*, 114(40):12811–12824, 2010. cited By (since 1996)16.

[21] K.A. Beauchamp, Y.-S. Lin, R. Das, and V.S. Pande. Are protein force fields getting better? a systematic benchmark on 524 diverse nmr measurements. *Journal of Chemical Theory and Computation*, 8(4):1409–1414, 2012. cited By (since 1996)21.

[22] D.-W. Li and R. Brüschweiler. Iterative optimization of molecular mechanics force fields from nmr data of full-length proteins. *Journal of Chemical Theory and Computation*, 7(6):1773–1782, 2011. cited By (since 1996)18.

[23] B.L. Foley, M.B. Tessier, and R.J. Woods. Carbohydrate force fields. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(4):652–697, 2012. cited By (since 1996)3.

[24] K.N. Kirschner, A.B. Yongye, S.M. Tschampel, J. González-Outeiriño, C.R. Daniels, B.L. Foley, and R.J. Woods. Glycam06: A generalizable biomolecular force field. carbohydrates. *Journal of Computational Chemistry*, 29(4):622–655, 2008. cited By (since 1996)270.

[25] E.I. Newhouse, D. Xu, P.R.L. Markwick, R.E. Amaro, H.C. Pao, K.J. Wu, M. Alam, J.A. McCammon, and W.W. Li. Mechanism of glycan receptor recognition and specificity switch for avian, swine, and human adapted influenza virus hemagglutinins: A molecular dynamics perspective. *Journal of the American Chemical Society*, 131(47):17430–17442, 2009. cited By (since 1996)20.

[26] C. Diehl, S. Genheden, K. Modig, U. Ryde, and M. Akke. Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *Journal of Biomolecular NMR*, 45(1-2):157–169, 2009. cited By (since 1996)30.

[27] L. Petersen, A. Ardèvol, C. Rovira, and P.J. Reilly. Mechanism of cellulose hydrolysis by inverting gh8 endoglucanases: A QM/MM metadynamics study. *Journal of Physical Chemistry B*, 113(20):7331–7339, 2009. cited By (since 1996)34.

[28] G. Hummer. Molecular binding: Under water's influence. *Nature Chemistry*, 2(11):906–907, 2010. cited By (since 1996)25.

[29] R. Baron, P. Setny, and J. Andrew McCammon. Water in cavity-ligand recognition. *Journal of the American Chemical Society*, 132(34):12091–12097, 2010. cited By (since 1996)65.

[30] P. Setny, R. Baron, and J.A. McCammon. How can hydrophobic association be enthalpy driven? *Journal of Chemical Theory and Computation*, 6(9):2866–2871, 2010. cited By (since 1996)47.

[31] R. Abel, T. Young, R. Farid, B.J. Berne, and R.A. Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008. cited By (since 1996)138.

[32] Z. Li and T. Lazaridis. Thermodynamic contributions of the ordered water molecule in hiv-1 protease. *Journal of the American Chemical Society*, 125(22):6636–6637, 2003. cited By (since 1996) 48.

[33] Z. Li and T. Lazaridis. The effect of water displacement on binding thermodynamics: Concanavalin A. *Journal of Physical Chemistry B*, 109(1):662–670, 2005. cited By (since 1996)57.

[34] S.D. Lella, M.A. Martí, R.M.S. Álvarez, D.A. Estrin, and J.C. Díaz Ricci. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *Journal of Physical Chemistry B*, 111(25):7360–7366, 2007. cited By (since 1996) 11.

[35] W. Humphrey, A. Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996. cited By (since 1996) 7861.

[36] D. R. Ferro and J. Hermans. A different best rigid-body molecular fit routine. *Acta Crystallographica Section A*, 33(2):345–347, Mar 1977.

[37] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 1. theory. *Journal of Physical Chemistry B*, 102(18):3531–3541, 1998. cited By (since 1996) 52.

[38] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids. *Journal of Physical Chemistry B*, 102(18):3542–3550, 1998. cited By (since 1996) 32.

[39] T. Lazaridis and M. Karplus. Orientational correlations and entropy in liquid water. *Journal of Chemical Physics*, 105(10):4294–4316, 1996. cited By (since 1996)73.

[40] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998. cited By (since 1996)4306.

[41] Francisco J. Solis and Roger J.B. Wets. Minimization by random search techniques. *Mathematics of Operations Research*, 6(1):19–30, 1981. cited By (since 1996)466.

[42] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1(2):227–235, 1992. cited By (since 1996)334.

[43] M. Marsili and J. Gasteiger. π charge distribution from molecular topology and π orbital electronegativity. *Croat. Chem. Acta*, 53:601–614, 1980. cited By (since 1996)100.

[44] J. Gasteiger and M. Marsili. *Tetrahedron*, 36:3210, 1980. cited By (since 1996)10.

[45] Pieter F.W. Stouten and Cornelius Frommel. Effective solvation term based on atomic occupancies fore use in protein simulations. *Molecular Simulation*, 10(2-6):97–120, 1993. cited By (since 1996)206.

[46] D.F. Gauto, S. Di Lella, C.M.A. Guardia, D.A. Estrin, and M.A. Martí. Carbohydrate-binding proteins: Dissecting ligand structures through solvent environment occupancy. *Journal of Physical Chemistry B*, 113(25):8717–8724, 2009. cited By (since 1996)13.

[47] D.F. Gauto, S. Di Lella, D.A. Estrin, H.L. Monaco, and M.A. Martí. Structural basis for ligand recognition in a mushroom lectin: Solvent structure as specificity predictor. *Carbohydrate Research*, 346(7):939–948, 2011. cited By (since 1996)3.

[48] J.M.J. Swanson, R.H. Henchman, and J.A. McCammon. Revisiting free energy calculations: A theoretical connection to mm/pbsa and direct calculation of the association free energy. *Biophysical Journal*, 86(1 I):67–74, 2004. cited By (since 1996)227.

[49] S. Genheden and U.L.F. Ryde. How to obtain statistically converged mm/gbsa results. *Journal of Computational Chemistry*, 31(4):837–846, 2010. cited By (since 1996)42.

[50] M.K. Gilson, M.E. Davis, B.A. Luty, and J.A. McCammon. Computation of electrostatic forces on solvated molecules using the poisson-boltzmann equation. *Journal of Physical Chemistry*, 97(14):3591–3600, 1993. cited By (since 1996)199.

[51] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995. cited By (since 1996)1824.

[52] Q. Lu and R. Luo. A poisson-boltzmann dynamics method with nonperiodic boundary condition. *Journal of Chemical Physics*, 119(21):11035–11047, 2003. cited By (since 1996)84.

[53] C.J. Cramer and D.G. Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chemical Reviews*, 99(8):2161–2200, 1999. cited By (since 1996)1348.

[54] W. Im, D. Beglov, and B. Roux. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation. *Computer Physics Communications*, 111(1-3):59–75, 1998. cited By (since 1996)248.

[55] D. Qiu, P.S. Shenkin, F.P. Hollinger, and W.C. Still. The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *Journal of Physical Chemistry A*, 101(16):3005–3014, 1997. cited By (since 1996)649.

[56] A. Ghosh, C.S. Rapp, and R.A. Friesner. Generalized born model based on a surface integral formulation. *Journal of Physical Chemistry B*, 102(52):10983–10990, 1998. cited By (since 1996)268.

[57] D. Bashford and D.A. Case. Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51:129–152, 2000. cited By (since 1996)595.

[58] M.S. Lee, F.R. Salsbury Jr., and C.L. Brooks III. Novel generalized born methods. *Journal of Chemical Physics*, 116(24):10606–10614, 2002. cited By (since 1996)252.

[59] A.N. Romanov, S.N. Jabin, Y.B. Martynov, A.V. Sulimov, F.V. Grigoriev, and V.B. Sulimov. Surface generalized born method: A simple, fast, and precise implicit solvent model beyond the coulomb approximation. *Journal of Physical Chemistry A*, 108(43):9323–9327, 2004. cited By (since 1996)20.

[60] L. David, R. Luo, and M.K. Gilson. Comparison of generalized born and poisson models: Energetics and dynamics of hiv protease. *Journal of Computational Chemistry*, 21(4):295–309, 2000. cited By (since 1996)73.

[61] V. Tsui and D.A. Case. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *Journal of the American Chemical Society*, 122(11):2489–2498, 2000. cited By (since 1996)286.

[62] N. Calimet, M. Schaefer, and T. Simonson. Protein molecular dynamics with the generalized born/ace solvent model. *Proteins: Structure, Function and Genetics*, 45(2):144–158, 2001. cited By (since 1996)74.

[63] T. Wang and R.C. Wade. Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. *Proteins: Structure, Function and Genetics*, 50(1):158–169, 2003. cited By (since 1996)51.

[64] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, 1990. cited By (since 1996)2210.

[65] T. Hou, J. Wang, Y. Li, and W. Wang. Assessing the performance of the molecular mechanics/poisson boltzmann surface area and molecular mechanics/generalized born surface area methods. ii. the accuracy of ranking poses generated from docking. *Journal of Computational Chemistry*, 32(5):866–877, 2011. cited By (since 1996)43.

Capítulo 3

Sitios de hidratación en la superficie de reconocimiento de carbohidratos

3.1. Introducción

Las simulaciones de dinámica molecular en solvente explícito surgen como una promisoria herramienta para el estudio de las interacciones agua-proteína. A través de está metodología, es posible alcanzar una descripción de la estructura y de las propiedades dinámicas de las moléculas de agua presentes en la superficie del sitio de reconocimiento carbohidratos (**CBS**). Además, utilizando las simulaciones de dinámica molecular en agua explícita, combinada con un análisis termoestadístico, las propiedades termodinámicas de las moléculas de agua presentes en el CBS pueden ser cálculadas y analizadas desde un punto de vista comparativo.[1,2]

Por ejemplo, Lazaridis et al., utilizando la teoría del fluido inhomogéneo, han sido capaces de calcular la contribución entálpica y entrópica a la energía libre de llevar una molécula de agua desde el seno del solvente hacia la superficie proteica para un grupo seleccionado de moléculas de agua presentes en el CBS.[2–6] Por otro lado, Abel et al., en base a los trabajos de Li et al., han estudiado la distribución de las moléculas de agua en el sitio de unión de ligandos del factor Xa, un importante blanco para el diseño de drogas involucradas en la vía de la coagulación, con el objetivo de poder diseñar una función semiempírica que discrimine aquellas drogas que se unen de las que no se unen al fXa.[1] Finalmente, Di Lella et al. han estudiado las propiedades de la estructura de solvatación sobre el CBS de Galectina 1 (Gal-1) a partir de simulaciones de dinámica molecular (DM), mostrando que las moléculas de agua presentes en el CBS juegan un rol activo en la unión proteína-ligando y que es posible predecir las regiones del espacio donde los átomos de oxígeno del ligando pueden ser encontradas en la estructura del complejo Gal-1/Galactosamina.[7]

En este contexto, podemos plantearnos los siguientes interrogantes:

- ¿Es posible describir la estructura de solvatación del CBS como un suma de regiones

del espacio cercanas a la superficie proteica en donde la probabilidad de encontrar moléculas de agua se mayor que en el seno del solvente?. En otras palabras, la superficie hidrofílica de la superficie proteica del CBS de lectinas, es capaz de retener moléculas de agua otorgándoles a las mismas una propiedad químico-física diferente a las del seno del solvente?

- ¿Se puede predecir qué moléculas de agua son desplazadas del CBS en el proceso de asociación entre las biomoléculas?. En ese caso, existe una preferencia respecto de qué grupo funcional del ligando las desplace?

- ¿Las propiedades termodinámicas de estas moléculas de agua, son similares o el entorno proteico marca diferencias significativas entre ellas?

- ¿Conocer los detalles estructurales y termodinámicos de la estructura de solvatación del CBS, tiene algún carácter predictivo respecto de poder predecir qué moléculas de agua son desplazadas hacia el seno del solvente por el ligando cuando se forma el complejo? Es decir, es posible utilizar esta información para guiar el diseño racional de drogas glicomiméticas.

Para abordar estás preguntas, y poder responderlas de manera general, se han seleccionado un conjunto de proteínas, principalmente lectinas, cuyas estructuras unidas a sus respectivos ligandos ha sido previamente resuelta mediante la técnica de difracción de rayos X y han sido depositadas en el *P*rotein Data Bank (**PDB**) y cuya elección fue motivada tanto por su relevancia fisiopatológica y por el hecho de que en todos los casos se espera la presencia de moléculas de agua fuertemente unidas a la superficie proteica juegue un rol clave en determinar la función de la mismas. Las proteínas seleccionadas son los dos módulos de la Sialidasa de bacteria multimodular (**CBM32** y **CBM40**) que unen Galactosa y Ácido siálico respectivamente, Galectina 3 (**Gal-3**) que une el disacárido N-Acetil-Lactosamina, Concanavalina-A (**Con-A**) que une un trimanósido y una proteína que une un decapéptido cíclico llamada Ciclofilina A (**Cyp-A**).

3.2. Ideas e hipótesis de trabajo

En la superficie proteica, se establecen principalmente interacciones entre los grupos dadores/aceptores de puentes de hidrógeno de los aminoácidos y las moléculas de agua circundantes que compiten con las interacciones $H_2O - H_2O$ del seno del solvente. En función de las características químico-físicas de las superficie proteica, que es inhomogénea desde el punto de vista topológico (es decir, no todos los grupos funcionales estarán expuestos al solvente de la misma manera), de distribución de cargas, de movilidad

intrínseca de sus residuos, etc., las moléculas de agua circundantes serán retenidas con distinto grado de afinidad por la proteína dependiendo de la zona de la misma con la que interactúen. Dada la similitud entre las interacciones de los grupos -OHs de los carbohidratos en los complejos biomoleculares, con la interacción de las moléculas de agua en la superficie proteica, es relevante plantearse como hipótesis de trabajo que aquellas moléculas de agua fuertemente retenidas en la superficie proteica, serán candidatas a ser desplazadas por los grupos hidroxilo del ligando. Dicho en otras palabras, cuánto más fuertemente retenidas estén las moléculas de agua por la proteína, mas desfavorecidas desde el punto de vista entrópico estarán y por ende, al momento de ser liberadas en el proceso de formación del complejo, contribuirán favorablemente a la energía libre de unión proteína-ligando.

3.3. Resultados y Discusión

Los resultados presentados en está sección están organizados de la siguiente manera: En primer lugar, se presentarán 5 secciones, una para cada proteína estudiada. Dentro de cada sección, se dará una breve descripción del CBS junto con la caracterización estructural y termodinámica de los WS determinados. Luego, los resultados serán interpretados en términos de su relación a las características del ligando. Por último, todos los datos serán discutidos en función a los aspectos generales de las propiedades estructurales y termodinámicas de los WS y su relación con la formación del complejo.

3.3.1. Caracterización estructural y termodinámica de los WS en el CBS de CBM32

El modulo de unión a carbohidratos 32 de la sialidasa de bacteria (CBM32), muestra una estructura de β-sandwich con un único CBS sobre uno de sus lados, mostrando preferencia por unir galactosa. Los principales residuos presentes en el CBS son Arg68, His37, Asn73 y Trp40 en donde se pueden detectar principalmente de 4 WS (ver figura 3.1 A) y cuyos parámetros termodinámicos se muestran en la tabla 3.1. Los resultados muestran que WS1 y WS4 tienen un alto valor de WFP, WS3 muestra un valor moderado y WS2 muestra un valor apenas mayor al del seno del solvente. Si se compara la posición de los WS con la posición de los átomos del ligando en el complejo (ver figura 3.1 B), es posible observar que el WS1 se encuentra sobre uno de los lados del anillo, a una distancia menor al contacto de van der Waals (menor a 2.5 Å) de los átomos de carbono del ligando, donde claramente se muestra que las aguas que pudieran ocupar este WS son desplazadas hacia el seno del solvente en el proceso de formación del complejo. Por otro lado, el WS3 y WS4 son reemplazados por el OH3 y el OH4 de la galactosa respectivamente, como se puede inferir a partir de los

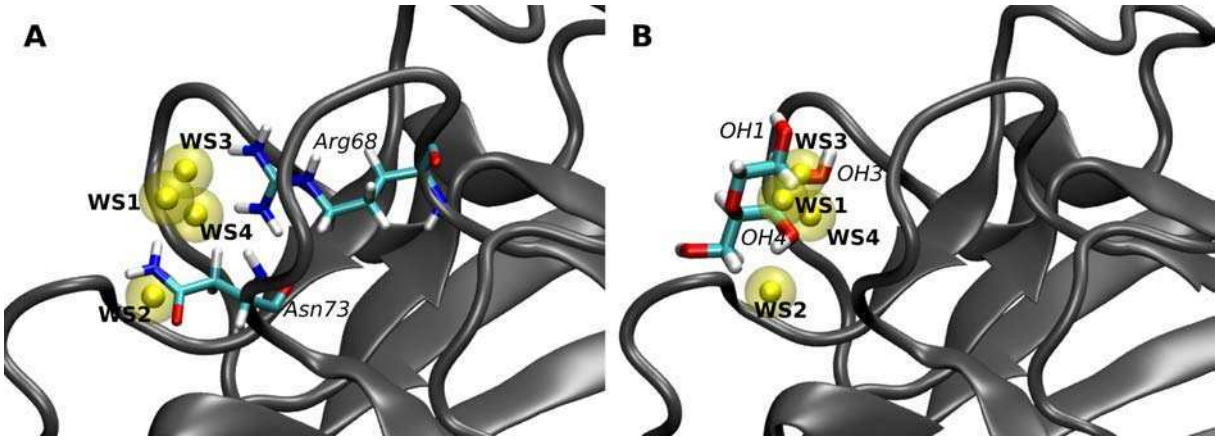

Figura 3.1. CBM32 con los WSs descriptos en la tabla 3.1 en esferas amarillas A) en la proteína libre y B) en el complejo con galactosa

Tabla 3.1. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de CBM32

WS	Átomo de ref.	Res.	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	WFP	R_{90}	R_{min}
1	ND2	Asn73	-9.0	-13.5	-22.5	-5.1	18.9	1.6	1.5
2	ND2	Asn73	-4.6	-16.4	-21.0	-3.6	2.4	4.4	2.5
3	NH1	Arg68	-13.6	-10.1	-23.7	-6.3	6.2	2.1	0.9
4	NH1	Arg68	-14.9	-9.1	-23.9	-6.5	11.5	1.7	1.2

* Los valores de energía, R_{90} y R_{min} están expresados en kcal/mol y Å, respectivamente.

bajos valores de R_{min}. Finalmente, el WS2 no se encuentra demasiado lejos del C6 de la galactosa, de modo tal que las aguas que visitan dicho WS, tienen bajas chances de ser desplazadas hacia el seno del solvente en el proceso de unión.

3.3.2. Caracterización estructural y termodinámica de los WS en el CBS de CBM40

El modulo de unión a carbohidratos 40 de la sialidasa de bacteria (CBM40), muestra un plegamiento β-sandwich con un único CBS que reconoce principalmente ácido siálico. Los principales residuos proteicos responsables del reconocimiento son Tyr70, Glu79, Arg81, Tyr150, Arg158 y Asn156. y el análisis a partir de los resultados de la simulación por dinámica molecular permitió identificar 7 WS con alto valor de WFP (ver figura 3.2 A) y cuyas propiedades termodinámicas se muestran en la tabla 3.2. En la misma, podemos observar que el WS1, WS3 y WS6 tienen los valores de WFP más altos, es decir que en promedio, es 15 veces más probable encontrar una molécula de agua en estos WS que en el seno del solvente. Si se compara la posición de los mismos con la posición de los átomos del ácido siálico en el complejo (ver figura 3.2 B), es posible observar que el WS1 definido a

Tabla 3.2. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de CBM40

WS	Átomo de ref.	Res.	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	WFP	R_{90}	R_{min}
1	ND2	Asn156	-6.7	-13.6	-20.3	-2.9	17.8	1.8	0.8
2	NH1	Arg151	-7.0	-12.2	-19.1	-1.7	8.5	3.1	1.7
3	NH2	Arg151	-11.5	-11.6	-23.0	-5.6	13.6	1.7	0.7
4	NH2	Arg81	-11.4	-12.7	-24.0	-6.6	7.5	2.1	1.4
5	OE1	Glu79	-12.4	-7.6	-20.0	-2.6	6.5	3.5	1.3
6	OH	Tyr158	-7.3	-14.3	-21.6	-4.2	18.8	1.7	0.8
7	ND1	His92	-9.1	-11.1	-20.2	-2.8	4.7	2.8	2.4

* Los valores de energía, R_{90} y R_{min} están expresados en kcal/mol y Å, respectivamente.

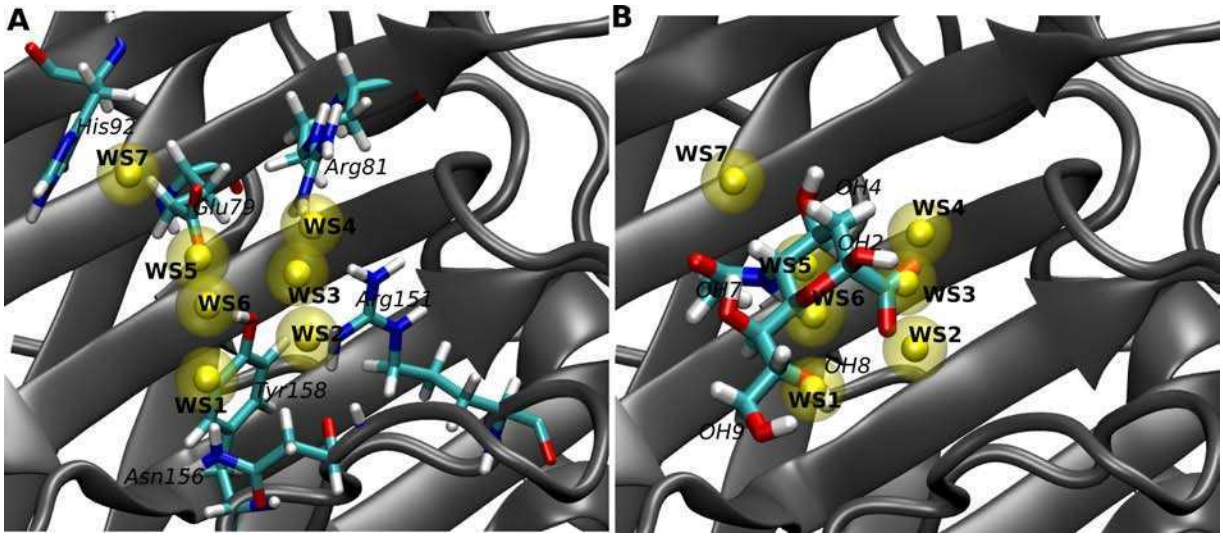

Figura 3.2. CBM40 con los WS descriptos en la tabla 3.2 en esferas amarillas A) en la proteína libre y B) en el complejo con ácido siálico

Tabla 3.3. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de carbohidratos (CBS) de Gal-3

WS	Atómo de ref.	Res.	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	WFP	R_{90}	R_{min}
1	NE2	His158	-10.1	-12.3	-22.4	-4.9	7.0	2.2	1.6
2	NH1	Arg162	-25.0	-3.5	-28.5	-11.4	9.1	1.5	0.4
3	NH1	Arg162	-11.0	-10.5	-21.5	-5.4	4.1	3.3	1.8
4	ND2	Asn174	-13.8	-10.4	-24.2	-6.4	18.5	1.5	0.3
5	OE2	Glu184	-19.0	-5.0	-24.0	-6.9	7.0	2.4	1.3
6	NH2	Arg186	-7.0	-14.7	-21.7	-4.8	6.6	4.1	2.3

* Los valores de energía, R_{90} y R_{min} están expresados en kcal/mol y Å, respectivamente.

partir de la interacción con el Asn156, es perfectamente reemplazado por el OH8 del ácido, el WS6 definido a partir de la Tyr158 es perfectamente reemplazado por el C6 del ácido siálico y finalmente, WS3 definido a partir de Arg151 es perfectamente reemplazado por el grupo ácido del ligando. En un segundo grupo, encontramos al WS5 definido por Glu79 el cual es reemplazado por el N del grupo N-acetilo. Luego el WS4 es también probablemente desplazado por el oxígenos perteneciente al grupo carboxilato, mostrando un R_{min} menor a 1.5 Å. Finalmente, las moléculas de agua que ocupan los WS2 y WS7 no parecen ser desplazados por algún átomo del ligando.

3.3.3. Caracterización estructural y termodinámica de los WS en el CBS de Gal-3

Gal-3 muestra la típica estructura de β-Sandwich con un único CBS que esta definido por los residuos 144 a 188 y es capaz de reconocer N-Acetil-Lactosamina. De los residuos pertenecientes al CBS, solo 9 fueron seleccionados (Arg 144, Asp148, His158, Glu165, Arg162, Asn174, Lys176, Glu184 y Arg186), que son los que presentan átomos potenciales para actuar como dadores o aceptores de puentes de hidrógeno con las moléculas de agua circundantes, y de esa manera establecer posibles WS. De esta manera, utilizando el protocolo que se describió en la sección métodos (capítulo 2), tomando como referencia los residuos del CBS seleccionados, se han podido establecer 6 WS en el CBS de Gal-3, con un valor de WFP mucho mayor que uno (ver figura 3.3 A) y cuyas propiedades termodinámicas se muestran en la tabla 3.3. Una vez determinados los WS, se han calculado las propiedades termodinámicas de los mismos, como se describe en el capitulo 2 y los resultados se muestran en la tabla 3.3. Como es posible observar a partir de la misma, la mayor parte de los WS identificados tiene al menos un valor de WFP mayor a 5, es decir, que es 5 veces más probable encontrar una molécula de agua dentro del espacio definido por el WS que en el seno del solvente. Por otro lado, los WS con el mayor valor de WFP (WS2 y WS4 con valores de WFP de 9.1 y 18.5 respectivamente) muestran valores muy bajos de R_{90}

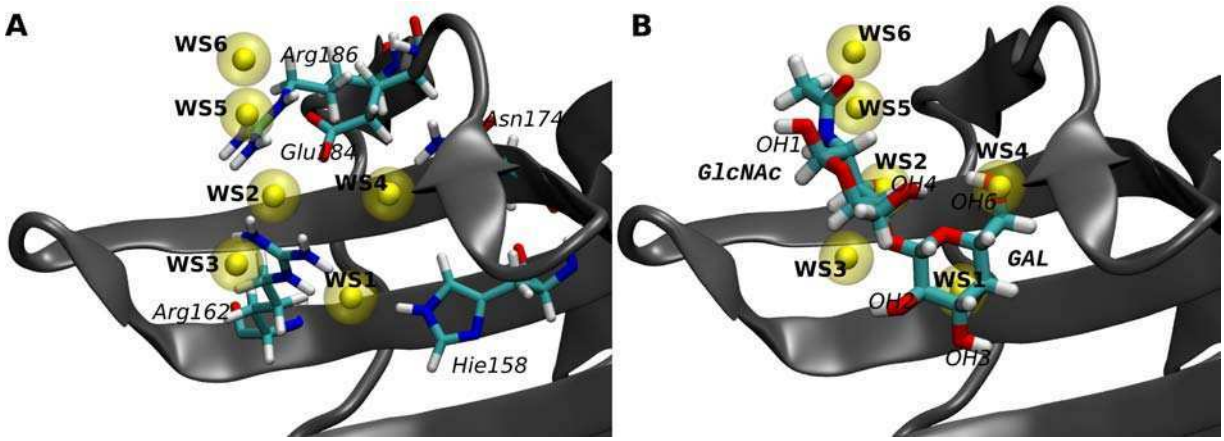

Figura 3.3. Gal-3 con los sitios de hidratación (WS) descriptos en la tabla 3.3 en esferas amarillas A) en la proteína libre y B) en el complejo con N-Acetil-Lactosamina

(menores a 1.5 Å). En un segundo grupo podemos encontrar a los WS1, WS5 y WS6, con valores de WFP alrededor de 7 y con valores de R_{90} mayores que WS2 y WS4.

Por otro lado, una medida cuantitativa del desplazamiento de aguas puede ser expresada en términos de la mínima distancia entre el centro de un determinado WS y el átomo del ligando distinto de hidrógeno más cercano al mismo, medido por el parámetro R_{min}, tal como se muestra en la tabla 3.3. Un análisis visual de la figura 3.3 B, muestra que los WS2 y WS4 son perfectamente reemplazados por los oxígenos del OH3 de GlcNAc y O6 del OH6 de la Galactosa (Gal) respectivamente. Además, las moléculas de agua que se encuentran en los sitios de hidratación en la proteína libre, muestran el mismo patrón de puentes de hidrógeno que los átomos de oxígeno del ligando. Para el WS1, si bien la coincidencia no es perfecta, este ocupa un espacio similar y los puentes de hidrógeno que establecen las moléculas de agua dentro del mismo son análogas a las establecidas por el átomo de oxígeno O4 de la galactosa. El WS5 por su parte, es reemplazado por el oxígeno carbonílico del grupo N-Acetilo y el WS6 es ocupado por el grupo metilo del mismo. Finalmente, las moléculas de agua que se encuentran en el WS3, el cual se define cerca del oxígeno de la unión glicosídica entre los dos monosacáridos, son desplazadas durante el proceso de unión por el mismo. En resumen, los resultados para Gal-3 muestran que la mayor parte de los WS son desplazados del CBS por el ligando y que cuanto mayor es el valor de WFP, mayor es la chance de ser ocupado por los átomos de oxígeno del carbohidrato.

3.3.4. Caracterización estructural y termodinámica de los WS en el CBS de Con-A

La estructura de Con-A está constituida principalmente por hojas β formando una estructura de β-sandwich en donde el CBS, que es capaz de reconocer un trimanósido, se

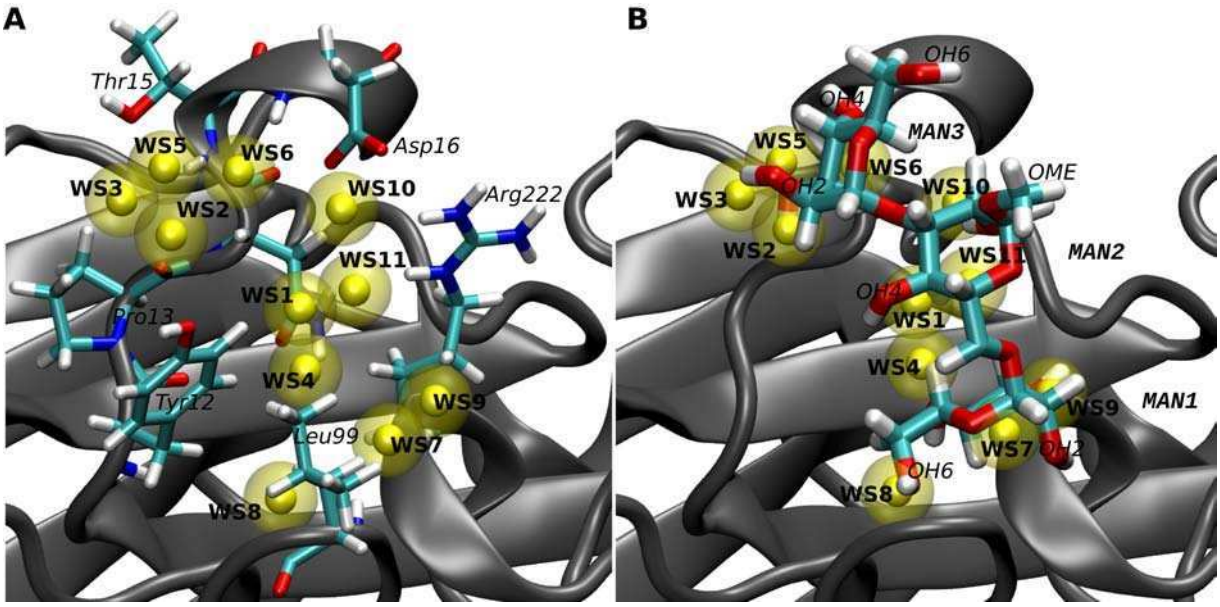

Figura 3.4. Con-A con los sitios de hidratación (WS) descriptos en la tabla 3.4 en esferas amarillas A) en la proteína libre y B) en el complejo con el trimanósido.

encuentra en el lado superior de la pequeña hoja beta. En el mismo, fue posible seleccionar siete residuos (Tyr12, Pro13, Asn14, Thr15, Asp16, Leu99 y Arg222) como aquellos capaces de establecer uniones puentes de hidrógeno con las moléculas de agua, pudiéndose identificar 11 WS (ver figura 3.4 A) y los parámetros termodinámicos calculados se muestran en la tabla 3.4. Como es posible observar en la misma, a excepción de los WS4, WS10 y WS11, todos los WS identificados tienen un alto valor de WFP. Finalmente, sobre la base de la estructura cristalográfica del trimanósido unido a Con-A (PDB ID 1ONA), es posible analizar que WS será reemplazado por algún átomo del ligando (ver figura 3.4 B). La primera manosa (MAN1) muestra varios de sus grupos hidroxilo estableciendo puentes de hidrógeno con la proteína; el OH4 forma interacciones puente de hidrógeno con el oxígeno carbonílico de Asp208. Luego, OH3, OH6 y O5 interaccionan con el N del grupo amida perteneciente Arg222 NH, Leu99NH y Tyr100NH respectivamente. De estos, OH6 es perfectamente reemplazado por WS8, O5 por WS7, en donde las moléculas de agua del WS, establecen las mismas interacciones puentes de hidrógeno con el *backbone* de la proteína que el ligando. Luego, el WS9 resulta perfectamente reemplazado por el OH3. Además, estos 3 WS muestran altos valores de probabilidad relativa. El OH2 de la manosa forma interacciones puente de hidrógeno que son ocupadas por las moléculas de agua que visitan el WS1 en la proteína libre en el transcurso de la dinámica molecular. También los puentes de hidrógeno que forma la Tyr12 y el O4 con el Asn14 son reemplazadas por el WS10 y WS11. Finalmente, el OH3 de la tercer manosa (MAN3) desplaza los WS2, WS3 y WS5. De estos, el WS2 y WS3 forman principalmente interacciones puente de hidrógeno con el O

Tabla 3.4. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de hidratos de carbono (CBS) de Concanavalina A

WS	Atómo de ref.	Res.	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	WFP	R_{90}	R_{min}
1	OH	Tyr12	-5.1	-15.1	-20.1	-2.7	7.7	2.9	1.5
2	OH	Tyr12	-5.9	-13.6	-19.5	-2.1	8.8	2.8	1.2
3	O	Pro13	-7.4	-13.9	-21.3	-2.3	12.8	1.9	1.4
4	ND2	Asn14	-14.6	-8.7	-23.3	-6.3	1.1	1.4	1.5
5	N	Thr15	-12.8	-8.4	-21.2	-3.7	7.2	3.9	0.9
6	OD1	Asp16	-9.5	-11.3	-20.8	-3.5	9.3	2.2	1.5
7	N	Leu99	-5.5	-15.9	-21.4	-4.0	9.8	3.1	1.1
8	N	Leu99	-10.7	-14.1	-24.8	-7.4	20.3	1.6	0.7
9	N	Arg222	-6.3	-15.1	-21.4	-4.0	16.9	2.1	0.6
10	ND2	Asn14	-16.9	-7.1	-24.0	-6.6	3.4	2.5	2.8
11	ND2	Asn14	-17.0	-5.6	-22.6	-5.2	3.1	1.8	1.9

* Los valores de energía, R_{90} y R_{min} están expresados en kcal/mol y Å, respectivamente.

carbonílico, mientras que el WS5 está formando interacciones puentes de hidrógeno con el grupo NH de la Thr15. El WS3 y el WS5 están muy cercanos entre sí, casi adyacentes unos a otros. El WS6, que esta formado principalmente por las interacciones dadas por el grupo carboxilo de la Asp16 y las moléculas de agua que visitan el mismo, es reemplazado por el OH4 de la manosa 3. Resulta también interesante destacar el WS4, el cual tiene un bajo valor de probabilidad relativa, que es posiblemente reemplazado por el carbono de esqueleto del primer manósido y muestra un bajo valor de R_{min} al C5 de MAN1.

3.3.5. Caracterización estructural y termodinámica de los WS en el CBS de Cyp-A

La Ciclofilina A (**Cyp-A**) es el blanco de un decapéptido cíclico conocido como Ciclosporina A (**CsA**), los cuales establecen numerosas interacciones hidrofóbicas y de tipo puente de hidrógeno entre sí. En el sito de unión de la CsA , se han podido detectar varios WS con alto valor de WFP (ver figura 3.5 A). Los parámetros termodinámicos calculados se muestran en la tabla 3.5. De manera análoga a como se procedió con las proteínas descriptas previamente, se comparó la relación existente entre la posición de los centros de masa de los WS en la proteína libre y la posición que ocupan los átomos del ligando en el complejo (ver figura 3.5 A). Los WS con más alto valor de WFP son el WS1, WS2, WS5 y WS6 y se observa que todos estos WS son desplazados por la CsA (ver figura 3.5 B). Por un lado, como es posible inferir a partir del valor de R_{min}, el WS1 es perfectamente ocupado por el oxígeno carbonílico del residuo 9 de CsA, estableciendo una unión puente de hidrógeno con el grupo NH del triptófano 121 (Trp121NH). Por otro lado, el WS2 es desplazado por el carbonilo del residuo 10 de CsA interactuando con la Arg55. El WS3 (que es formado

Tabla 3.5. Parámetros termodinámicos y estructurales para los sitios de hidratación (WS) en las proximidades del sitio de reconocimiento de ligandos de CiclofilinaA

WS	Átomo de ref.	Res.	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	WFP	R_{90}	R_{min}
1	NE1	Trp121	-5.1	-15.9	-20.8	-3.4	11.0	2.7	0.4
2	NH2	Arg55	-13.1	-11.0	-24.3	-6.9	11.5	1.8	0.9
3	NE2	Gln63	-6.4	-15.6	-21.9	-4.5	3.3	2.7	1.0
4	OE1	Gln111	-7.3	-12.4	-19.7	-2.3	6.0	2.4	1.8
5	O	Asn102	-11.0	-9.1	-20.1	-2.7	10.1	2.4	1.2
6	O	Asn102	-7.0	-12.0	-19.0	-1.6	8.8	2,7	1.1

* Los valores de energía, R_{90} y R_{min} están expresados en kcal/mol y Å, respectivamente.

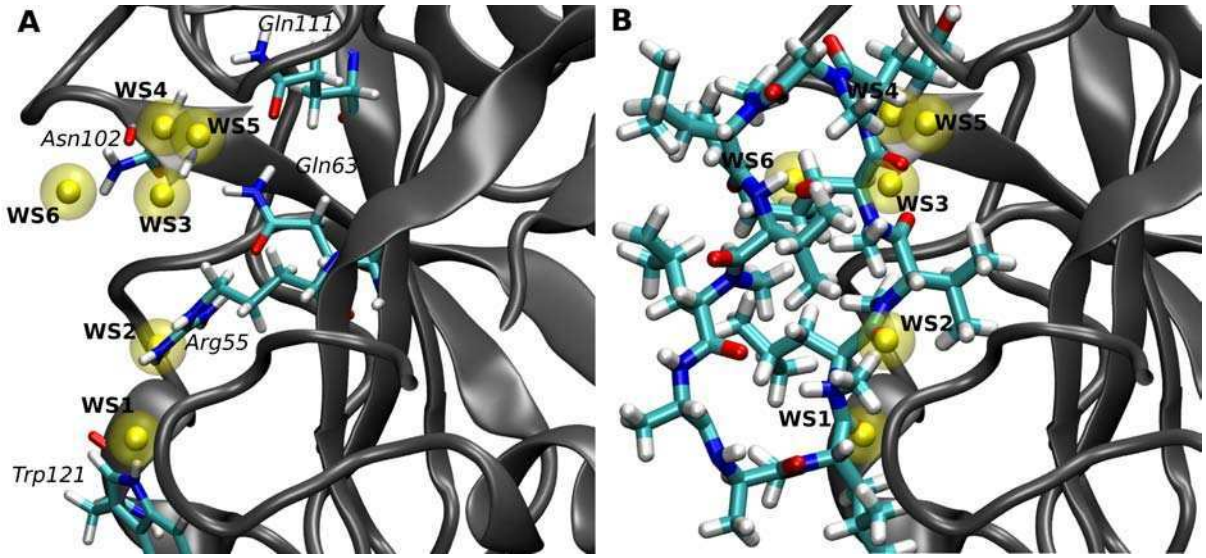

Figura 3.5. Ciclofilina A con los WS descriptos en la tabla 3.5 en esferas amarillas A) en la proteína libre y B) en el complejo

por las interacciones puentes de hidrógeno establecidas entre las moléculas de agua que visitan el sitio y el NE2 de Gln63) es reemplazado por el O carbonílico de residuo 1 de CsA. Luego,el WS4 tiene también un alto valor de WFP y es probablemente desplazado por el carbonilo del residuo 1. El WS6 es desplazado por el O carbonílico del residuo 2 y por último, las moléculas de agua que ocupan el WS5 son desplazadas por la cadena lateral hidrofóbica del residuo 1 de CsA.

3.4. Discusión

Tomando el conjunto de los WS determinados y sus propiedades termodinámicas, es posible plantear un análisis global de los mismos. La tabla 3.6 muestra los valores estadísticos promedio para todos los WS, pudiéndose observar a partir de la misma que aquellas

Tabla 3.6. Promedio, desvió estándar (s), valores mínimo y máximo para todos los parámetros termodinámicos y estructurales calculados

	$< \mathbf{E_p} >$	$< \mathbf{E_w} >$	$< \mathbf{E_T} >$	$< \mathbf{\Delta E} >$	**WFP**	$\mathbf{R_{90}}$	$\mathbf{R_{min}}$
Media	-10.6	-11.1	-21.5	-4.2	8.8	2.4	1.3
s	5.0	3.6	2.6	2.6	5.1	0.8	0.6
Mín.	-25.0	-16.4	-28.5	-11.4	2.3	1.4	0.3
Máx.	-3.5	-2.0	-14.3	-1.6	20.3	4.4	2.8

* Los valores de energía, R_{90} y R_{min} están expresados en Kcal/mol y Å, respectivamente.

moléculas de agua que visitan los mismos se estabilizan desde el punto de vista energético ($< \Delta E >$ <0). Otro aspecto interesante se observa en el hecho de que $< E_T >$ muestra una menor dispersión que sus dos componentes ($< E_p >$ y $< E_w >$), como se refleja en los valores de desvío estándar (s) y min/máx. Esto es razonable si tenemos en cuenta que las moléculas de agua no pueden establecer más de 4 interacciones puente de hidrógeno con su entorno. Consecuentemente, se establecerán un mayor número de interacciones con la proteína en detrimento de las interacciones agua-agua. Esto puede ser semi-cuantitativamente estimado a partir del coeficiente de correlación entre $< E_p >$ y $< E_w >$ de -0.77. Este hecho limita la cantidad de energía que una molécula de agua puede liberar cuando la misma se asocia a la superficie proteica, la cual parece ser en este caso de -11.4 Kcal/mol. Los datos presentados aquí permiten también determinar cuál es el parámetro que mejor predice la posición del ligando en el complejo proteína-ligando, al menos desde un punto de vista comparativo y cualitativo. En este sentido, si asumimos que R_{min} es un buen estimador de cuán bien un WS refleja la ocupancia de un átomo del ligando en el complejo proteína-ligando, es posible correlacionarlo con los parámetros termodinámicos calculados. Para esto, resulta interesante calcular el coeficiente de correlación para cada uno de los parámetros energéticos, $< E_p >$, $< E_w >$ o $< E_T >$ con R_{min} en donde los valores de correlación están entre 0.20 y 0.35, indicando una pobre correlación. Mientras que si correlacionamos WFP o R_{90} con R_{min} se encuentra que los valores de correlación son 0.55 y 0.61, respectivamente. Aunque estos valores del coeficiente de Pearson están lejos de ser perfectos, muestran que **tanto WFP como R_{90}, que contienen contribuciones entrópicas del solvente (lo cual no sucede con la energía de interacción), son mejores predictores de qué moléculas de agua tienen mayores chances de ser desplazadas en el proceso de formación del complejo proteína-carbohidrato**.

3.5. Conclusión

Utilizando simulaciones de dinámica molecular en solvente explícito, es posible definir regiones del espacio adyacentes a las zonas hidrofílicas de la superficie proteica en donde la probabilidad de encontrar moléculas de agua es alta (mayor que en el seno del solvente). De esta modo, se han podido analizar las propiedades termodinámicas y estructurales de más de 40 sitios de hidratación para seis proteínas distintas, Gal-1,[7] Gal-3, Con-A, Cyp-A, CBM32 y CBM40. Los resultados muestran que la probabilidad de encontrar moléculas de agua dentro de un sitio de hidratación, medida a través de WFP, está directamente relacionada con la chance de encontrar un átomo pesado del ligando (principalmente oxígeno) en el complejo proteína-ligando.

Esta información puede ser utilizada para analizar en detalle la estructura de solvatación del CBS y su conexión con los posibles complejos proteína-carbohidrato para poder sugerir adiciones de grupos -OH que desplacen moléculas de agua presentes en los WS con el fin de incrementar la afinidad de drogas, en especial aquellas glicomiméticas.

3.6. Metodología específica del capítulo

Las estructuras de los complejos proteína-carbohidrato fueron tomadas del *Protein Data Bank* (PDB), con los siguientes códigos (PDB ID): 1ONA para concanavalina A,[8] 1a3k para Galectina3,[9] 1mik para Ciclofilina A[10] y 2v73 y 2v72 para los módulos CBM32 y CBM40[11] respectivamente. Para cada complejo, los ligandos fueron borrados y solo se han simulado las estructuras proteicas como monómeros, con el objetivo de reducir el tiempo de cómputo. Dado que los átomos de hidrógeno no son resueltos en las estructuras cristalográficas, los mismos fueron agregados utilizando el modulo LEAP del programa Amber10.[12] Para aquellos aminoácidos con más de un posible estado de protonación (ácido aspartico (Asp), ácido glútamico (Glu), lisina (Lys), y arginina (Arg)) les fue asignado el estado de protonación correspondiente al pH fisiológico. En cuanto a la histidina, el estado de protonación asignado fue aquel que favoreciera la formación de puentes de hidrógeno en la estructura cristalina y luego cada proteína fue inmersa en una caja octaédrica truncada de moléculas de agua (solvente explícito) de tres centros (modelo de agua TIP3P). Cada sistema fue primero optimizado utilizando el algoritmo del gradiente conjugado para 2000 pasos, seguido por un paso de termalización de 200 ps. de dinámica molecular a volumen constante (en este lapso, la temperatura del sistema se incremento lenta y gradualmente desde 100 hasta 300 K). El calentamiento fue seguido por otros 200 ps. de dinámica molecular a presión constante con el objetivo de equilibrar la densidad del sistema. Durante este proceso, los átomos Cα pertenecientes al *backbone* de la proteína

fueron restringidos por un un potencial armónico de 1 kcal/mol. Luego del proceso de equilibración, se generaron 20 ns. de dinámica molecular a presión y temperatura constante para cada proteína (durante la producción de las trayectorias de dinámica molecular, no se aplico un potencial armónico restrictivo sobre los átomos de la proteína). Los parámetros para todos los aminoácidos corresponden al campo de fuerzas de Amber ff99SB.[13] La temperatura y la presión se mantuvieron constantes utilizando el termostato y el barostato de Berendsen,[14] respectivamente. Todas las simulaciones fueron hechas en condiciones periódicas de contorno, utilizando el método de las sumas de Ewald para las interacciones electrostáticas de largo alcance.[15] Por último, el algoritmo SHAKE fue aplicado a todos las uniones covalentes que tengan hidrógeno, de manera que sea posible utilizar una ventana temporal de 2 fs.

Referencias

[1] Robert Abel, Tom Young, Ramy Farid, Bruce J. Berne, and Richard A. Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008.

[2] Z. Li and T. Lazaridis. Thermodynamics of buried water clusters at a protein-ligand binding interface. *Journal of Physical Chemistry B*, 110(3):1464–1475, 2006. cited By (since 1996) 34.

[3] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 1. theory. *Journal of Physical Chemistry B*, 102(18):3531–3541, 1998. cited By (since 1996) 52.

[4] T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids. *Journal of Physical Chemistry B*, 102(18):3542–3550, 1998. cited By (since 1996) 32.

[5] Z. Li and T. Lazaridis. Thermodynamic contributions of the ordered water molecule in hiv-1 protease. *Journal of the American Chemical Society*, 125(22):6636–6637, 2003. cited By (since 1996) 48.

[6] Z. Li and T. Lazaridis. The effect of water displacement on binding thermodynamics: Concanavalin a. *Journal of Physical Chemistry B*, 109(1):662–670, 2005. cited By (since 1996) 53.

[7] S.D. Lella, M.A. Martí, R.M.S. Álvarez, D.A. Estrin, and J.C. Díaz Ricci. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *Journal of Physical Chemistry B*, 111(25):7360–7366, 2007. cited By (since 1996) 11.

[8] R. Loris, D. Maes, F. Poortmans, L. Wyns, and J. Bouckaert. A structure of the complex between concanavalin a and methyl-3,6-di-o- (α-d-mannopyranosyl)-α-d-mannopyranoside reveals two binding modes. *Journal of Biological Chemistry*, 271(48):30614–30618, 1996. cited By (since 1996) 53.

[9] J. Seetharaman, A. Kfanigsberg, R. Slaaby, H. Leffler, S.H. Barondes, and J.M. Rini. X-ray crystal structure of the human galectin-3 carbohydrate recognition domain at 2.1-Å resolution. *Journal of Biological Chemistry*, 273(21):13047–13052, 1998. cited By (since 1996) 187.

[10] V. Mikol, C. Papageorgiou, and X. Borer. The role of water molecules in the structure-based design of (5-hydroxynorvaline)-2-cyclosporin: Synthesis, biological activity, and crystallographic analysis with cyclophilin a. *Journal of Medicinal Chemistry*, 38(17):3361–3367, 1995. cited By (since 1996) 46.

[11] A.B. Boraston, E. Ficko-Blean, and M. Healey. Carbohydrate recognition by a large sialidase toxin from clostridium perfringens. *Biochemistry*, 46(40):11352–11360, 2007. cited By (since 1996) 35.

[12] D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005. cited By (since 1996) 1725.

[13] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz Jr., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995. cited By (since 1996) 6802.

[14] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, A. Dinola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984. cited By (since 1996) 9196.

[15] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An n·log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. cited By (since 1996) 5297.

Capítulo 4

Rol del solvente en la selectividad epimérica

4.1. Introducción

La lectina que se encuentra en el hongo *Agaricus bisporus*, (**ABL** por sus siglas en inglés Agaricus Bisporus Lectin) tiene una particular propiedad vinculada a unir selectivamente y con alta afinidad al antígeno de Thomsen-Friedenreich o antígeno-T el cual, es un disacárido (Galβ1-3GalNAc) unido tanto a serina como a treonina de glicoproteínas sobre la superficie celular y se encuentra oculto en células sanas, mientras que se encuentra expuesto en un alto porcentaje de células de carcinoma humano y otros tejidos neoplásicos.[1,2] ABL surge entonces como una importante proteína terapéutica, dada su capacidad de inhibir de manera reversible la proliferación de células epiteliales malignas de cultivo sin una aparente toxicidad sobre las células normales, debido al bloqueo selectivo de la secuencia de localización nuclear de la proteína importadora Orp-150, que expresa la forma sialilada del ligando de ABL (sialil-2,3-galactosil-β1,3-N-Acetilgalactosamina), la cual es esencial para el funcionamiento celular.[3,4] Luego de la dilucidación estructural de ABL por cristalografía de rayos-X, se pudo caracterizar el sitio de unión del antígeno-T, junto con la presencia de dos sitios de unión a carbohidratos (**CBS**, por sus siglas en inglés "Carbohydrate Binding Site") que son independientes entre sí y tienen la capacidad de distinguir entre dos monosacáridos que difieren únicamente en la configuración del grupo hidroxilo unido al C4, axial o ecuatorial (N-Acetil-Galactosamina y N-Acetil-Glucosamina respectivamente), es decir, que son epímeros entre sí (ver figura 4.1 y 4.2).[1]

Aunque la primera aproximación para comprender la interacción proteína-carbohidrato es abordada en principio por experimentos de rayos-X y/o RMN, como ya mostramos previamente, comprender el detalle fino de la regulación y discriminación, requiere de la caracterización y análisis de la dinámica del solvente y de la las macromoléculas involucradas, que puede ser aportado por las simulaciones de dinámica molecular.

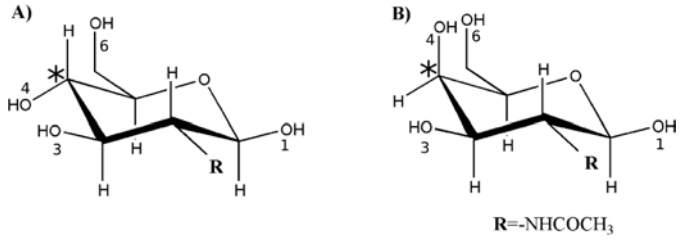

Figura 4.1. Representación esquemática de **A)** β-N-Acetil-D-Glucosamina (GlcNAc) y **B)** β-N-Acetil-D-Galactosamina (GalNAc). El átomo de carbono que está unido al O4 ecuatorial o axial respectivamente, está marcado con * .

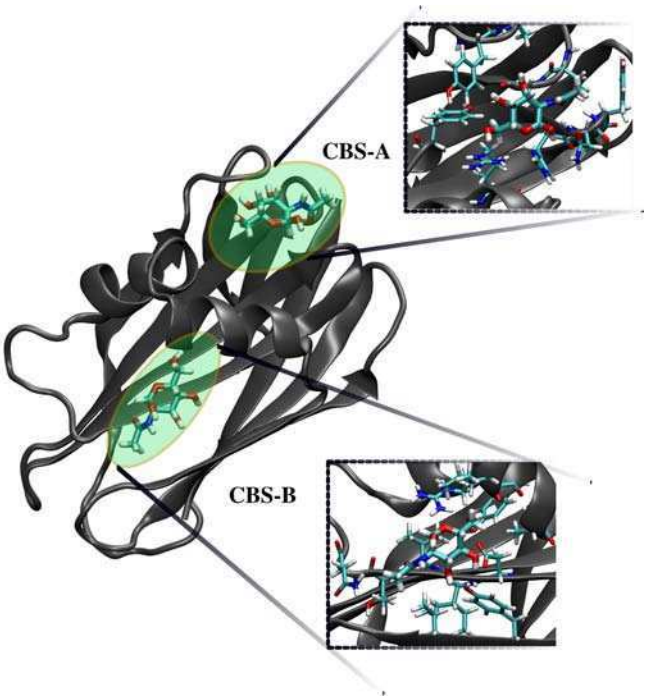

Figura 4.2. Representación del dominio de unión de carbohidratos (CRD) de ABL, donde se muestran los dos sitios de unión a carbohidratos (CBS-A y CBS-B) unidos a β-N-Acetil-D-Galactosamina (GalNAc) y β-N-Acetil-D-Glucosamina (GlcNAc), respectivamente.

4.2. Ideas e hipótesis

La inusual selectividad epimérica que se infiere a partir de la estructura de rayos-X de ABL, se presenta como un caso interesante para poder dilucidar las sutilezas de la selectividad y discriminación de ligandos, en el presente caso entre epímeros, por parte de sendos CBS y plantear como hipótesis inicial que el rol del solvente es clave en el proceso de reconocimiento. De este modo, las herramientas provistas por los métodos computacionales surgen como un complemento invaluable para comprender los resultados que se observan a partir de la estructura de rayos-X de los complejos, ya que nos permiten crear *I*n-silico aquellos complejos que no es posible obtener experimentalmente, y comparar las diferencias, en base a la información dinámica provista por los resultados de simulación de dinámica molecular de los mismos con el que se entiende como más estable, en base a los resultado experimentales. Por otro lado, como ya mostramos, es posible estudiar la estructura del solvente en el receptor libre, y poder correlacionar esta información con la estructura de ambos complejos (el observado experimentalmente y el que no), bajo la hipótesis alcanzada en el capítulo previo, vinculada a que aquellas moléculas de agua que se encuentran retenidas en el CBS de la proteína libre, pueden actuar como predictoras de la posición que los grupos -OHs del carbohidrato irán a ocupar en el complejo.

4.3. Resultados

Los resultados son analizados independientemente para cada CBS de ABL. En cada caso, se ha hecho un análisis estructural y dinámico de puentes de hidrógeno, un análisis energético y el estudio de la estructura de solvatación de los CBS. Luego se vincularán estos resultados con la especificidad de unión. Las dinámicas globales para todas las simulaciones son estables y se observa que alrededor 20 ns de simulación son suficientes para alcanzar el equilibrio.

4.3.1. Análisis estructural y dinámico del CBS-A de ABL

A partir del análisis de la estructura cristalográfica, es posible observar que el CBS-A de ABL está principalmente determinado por los residuos Arg107, Tyr98, Tyr28, Gly49, Ser48, Tyr74, Asn73 y His72 mientras que la simulación por dinámica molecular muestra, como es esperado, que el ligando se mantiene estable y unido a la proteína durante en el intervalo de tiempo simulado (el valor promedio de RMSD de los átomos de C del ligando con respecto a la estructura de referencia cristalográfica es de 1.2 Å), especialmente cuando la His72 es simulada en su estado tautomérico **HIE**, es decir con la histidina protonada en el Nϵ. Para la histidina en estado **HID** (la histidina protonada en el Nδ) el ligando parece

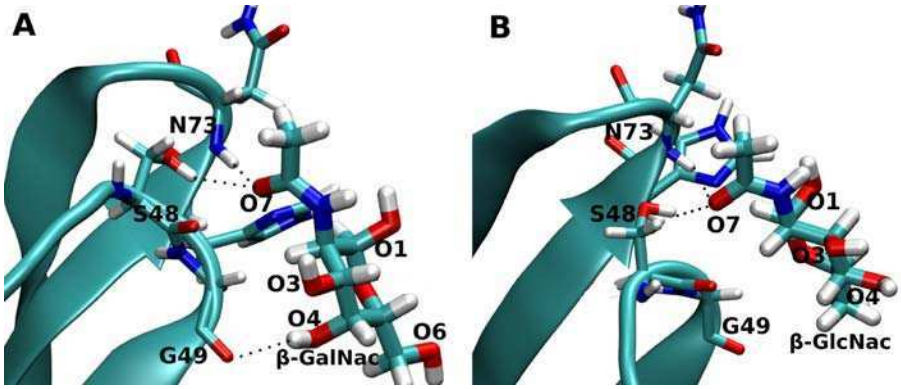

Figura 4.3. Estructura representativa de **A)** β-N-Acetil-D-Galactosamina (GalNAc) y **B)** β-N-Acetil-D-Glucosamina (GlcNAc) unidos al CBS-A de ABL

unirse más débilmente, como es posible evidenciar en base mayor movilidad observada en la dinámica molecular del complejo. En la figura 4.3 A, se observa una representación de la estructura de GalNAc unido al CBS-A de ABL y en la tabla 4.1 los correspondientes valores de análisis de puentes de hidrógeno. En el estado HIE, se establece una fuerte interacción puente de hidrógeno (>90 % de ocupancia) entre el grupo OH de la Ser48 y O carbonílico del grupo Acetamida de GalNAc (GalNAcO7). También se establecen puentes de hidrógeno fuertes entre el NH del *backbone* perteneciente a la Asn 73 y el O de la Acetamida y entre el O carbonílico de Gly 49 y el OH4 de GalNac (GalNAcO4). Finalmente, una interacción puente de hidrógeno débil, aquellas con un porcentaje de ocupancia menor al 10 %, puede ser establecido entre Asn73OD y GalNAcO1. Cuando la His72 es cambiada al tautómero HID, el ligando se desplaza y ahora se establecen 3 interacciones puente de hidrógeno pero más lábiles. No obstante, nuevas interacciones puente de hidrógeno transientes son establecidas entre Ser 48-O, Gly 49-O y GalNAcO3 y O4. Por otro lado, los resultados (en particular para el estado tautomérico HIE) son consistentes con aquellos observados por cristalografía de rayos-X, en los cuales los principales puentes de hidrógeno entre el ligando y el receptor se establecen entre Asn73-N y Ser48-OG con GalNAcO7 y el O carbonilico de Gly49 con el O3 y O4 de GalNAc. Estos resultados confirman que la histidina protonada en el Nϵ, estado tautomérico HIE, establece interacciones puente de hidrógeno más estables respecto del estado HID, y por lo tanto se espera que sea este el estado de protonación que adopta la misma en el complejo.

Si el ligando ahora es reemplazado por su epímero (GlcNAc), como se muestra en la figura 4.3 B y la tabla 4.1, algunas interacciones se mantienen mientras que otras, se debilitan de manera significativa (en particular las interacciones que se establecen entre el grupo hidroxilo ecuatorial O4 y el oxígeno carbonílico de Gly49O). Una inspección visual de la dinámica de los complejos, muestra que el ligando GlcNAc se mantiene unido de manera

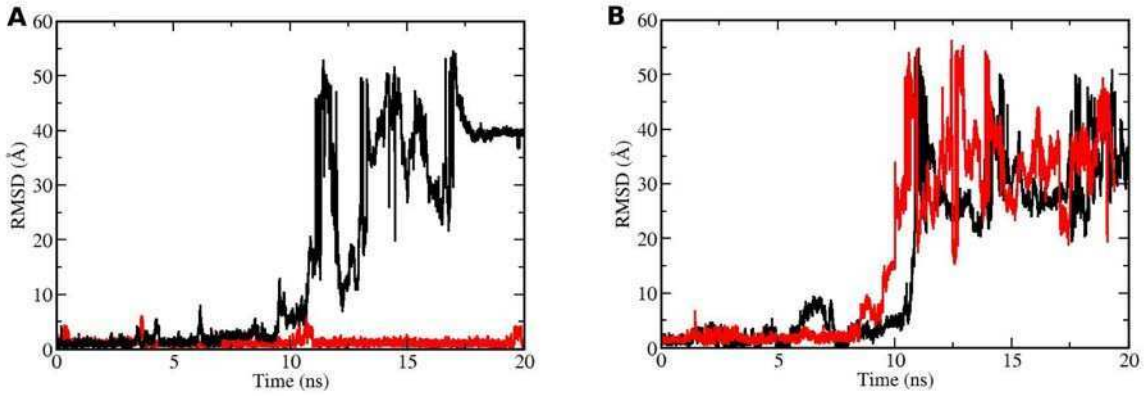

Figura 4.4. RMSD de los átomos de carbono de los carbohidratos (GalNAc en rojo y GlcNAc en negro) en el CBS-A vs. el tiempo de simulación. Los diferentes estados tautoméricos de las His72 fueron probados: **A)** HIE (histidina protonada en el Nϵ) y **B)** HID (histidina protonada en el Nδ).

transiente al CBS-A y se libera del mismo alrededor de los 10-15 ns. de simulación, como se evidencia en la figura 4.4. En la misma podemos observar el gráfico de RMSD de los átomos de carbono de los carbohidratos vs. tiempo, utilizando la estructura cristalográfica como referencia, variando tanto el estado tautomérico de la His72, como el ligando que ocupa el CBS-A (GalNAc o GlcNAc) y se muestra claramente que el GlcNAc, ligando invertido, se mantiene unido en los primeros 5 ns. y comienza a disociarse entre los 5 y 10 ns., mientras que luego de ese lapso es completamente liberado del CBS-A (ver línea negra en la figura 4.4 A y B). El análisis de puentes hidrógeno, realizado para la primera parte de la simulación en donde el ligando se mantiene en el CBS (durante los primeros 10 ns.) muestra que para el sistema HIE, las interacciones puentes de hidrógeno entre la Ser48 OH y GlcNAcO7 y Asn73NH y GlcNAcO7 se preservan con similar ocupancia y parámetros estructurales que el ligado natural. Sin embargo, las interacciones entre el O carbonílico de la Gly49 y el O4 ecuatorial GlcNAcO4 se pierde completamente. Cuando la His72 es simulada en el estado tautomerico HID, las interacciones puente de hidrógeno entre la Ser48 y Asn73 se pierden completamente y se establecen de manera temporaria interacciones entre los O carbonílico de Gly49, Ser48 con GlcNAcO3 y O4. De este modo se evidencia que el CBS-A de ABL puede unir GalNAc y no su epímero GlcNAc.

Para ganar mayor claridad en el mecanismo de unión y discriminación, se ha hecho un análisis termodinámico (como se describe en el capítulo 2) para estimar la energía unión de GalNAc al CBS-A con His72 en sus dos estados tautoméricos. Dado que no se establece una interacción estable entre el complejo de GlcNAc y el CBS-A, el análisis no ha podido realizarse sobre el ligando invertido. Los resultados presentados en la tabla 4.2, confirman la sutil preferencia observada para la His72 protonada en el Nϵ (HIE).

Tabla 4.1. Porcentajes de ocupancia de puentes de hidrógeno del CBS-A

Aceptor	Donor	GalNAc-HIE	GalNAc-HID	GlcNAc-HIE	GlcNAc-HID
GalNAc/GlcNAc O7	Ser48 HO	91.6	35.7	97.6	<1.0
GalNAc/GlcNAc O7	Asn 73 HN	50.0	35.4	35.0	<1.0
Gly 49 O	GalNAc HO4	42.7	13.1	<1.0	33.1
Asn 73 OD1	GalNAc/GlcNAc HO1	6.5	<1.0	18.2	<1.0
Ser 48 O	GalNAc HO4	<1.0	6.6	<1.0	6.8

Tabla 4.2. Valores medios para la energía de interacción calculada utilizando MM-GB(SA)

	CBS-A: GalNAc-HIE	CBS-A: GalNAc-HID
Ele	-19.1	-18.9
vdW	-18.3	-17.3
GAS	-37.4	-36.1
ΔG SV	26.9	27.0
GB_{TOT}	-10.5	-9.5

* Los valores de energía están expresados en kcal/mol.

Es interesante resaltar que las contribuciones electrostáticas y de vdW son de la misma magnitud para la interacción de GalNAc con el CBS-A. Por otro lado, para GalNAc-HIE, el OH axial contribuye favorablemente con una energía de 7.73 kcal/mol y para GalNAc-HID esta es una contribución favorable con una energía de unión de 6.41 kcal/mol. Estos valores corresponden a alrededor de la tercera parte del total de la energía electrostática total mostrando el rol clave jugado por esta región de la interacción proteína-ligando en modular el proceso global de unión y discriminación epimérica.

4.3.2. Análisis estructural y dinámico del CBS-B de ABL

El CBS-B de ABL, cuyo ligando natural es GlcNAc, está formado por los residuos Tyr114, Asp79, Ile80, Thr82, Asn83, Val81, Arg103, Ile102 y Tyr97. La simulación por dinámica molecular muestra que el ligando natural GlcNAc establece numerosas interacciones puente de hidrógeno con la proteína, que estabilizan al mismo en el CBS. Además, la posición y la orientación del ligando no difieren significativamente de la estructura cristalográfica (el valor de RMSD de la estructura promedio de los átomos de C respecto la estructura cristalográfica inicial es de 0.6 Å) sugiriendo que las interacciones son muy estables. La estructura representativa del complejo se muestra en la figura 4.5 A y el análisis de puentes de hidrógeno se resume en la tabla 4.3 en donde se muestra que GlcNAc se une al CBS-B de ABL estableciendo numerosas interacciones puente de hidrógeno. Asp79 forma dos uniones puente de hidrógeno muy fuertes (>98 % de ocupancia) con el oxígeno ecuatorial de GlcNAcO4 y GlcNAcO6. Por otro lado, Thr82NH establece una unión puente de hidrógeno fuerte con GlcNAcO7 e interacciones solo de manera transiente con el O4 de GlcNAc. Si el

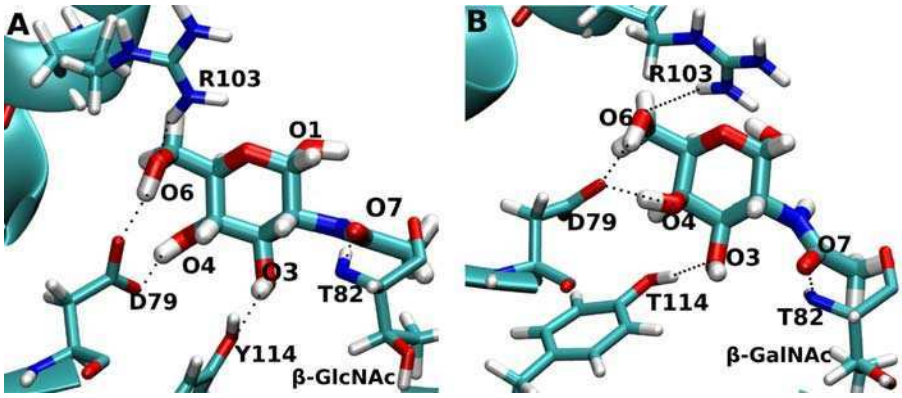

Figura 4.5. Estructura representativa de **A)** β-N-Acetil-D-Glucosamina (GlcNAc) y **B)** β-N-Acetil-D-Galactosamina (GalNAc) unidos al CBS-A de ABL

Tabla 4.3. Porcentajes de ocupancia de puentes de hidrógeno del CBS-B

Aceptor	Donor	GlcNAc	GalNAc
Asp 79 OD2	GlcNAc/GalNAc HO4	99.9	21.8
Asp 79 OD1	GalNAc/GalNAc HO4	<1.0	91.0
Asp 79 OD1	GlcNAc/GalNAc HO6	99.4	77.2
GlcNAc/GalNAcO7	Thr 82 NH	79.0	69.6
GlcNAc/GalNAcO3	Tyr 114 HO	72.5	93.6

ligando es cambiado por su epímero a GalNAc, el patrón de interacciones se mantiene casi inalterado pero algunas muestran una ocupancia más baja, en especial las interacciones entre el grupo carboxilato de Asp79 con GalNAcO4 y GalNAcO6, y Thr82NH con GalNAcO7. Es notable destacar la dinámica de Asp79. Cuando el ligando es GlcNAc, el grupo carboxilato de Asp79 se mantiene fijo y solamente un átomo de oxígeno forma puente de hidrógeno con el HO4 durante toda la simulación. Por otro lado, cuando se simula con el ligando invertido (GalNAc), el grupo carboxilato rota alternando la interacción entre los dos O del grupo carboxilato y el OH4 del azúcar.

Del mismo modo que para el CBS-A, se calcularon las contribuciones energéticas de la unión de ligando al CBS-B para GalNAc y GlcNAc utilizando el método de solvente continuo como se muestra en la tabla 4.4. Los resultados muestran una clara preferencia para el ligando natural (alrededor de 8.0 kcal/mol). El análisis de las contribuciones individuales muestran, de manera consistente con el análisis de la interacciones puente de hidrógeno, que la preferencia recae principalmente en la componente electrostática. También se cálculo la contribución de OH4 a la energía de unión mostrando que para GlcNAc, el OH ecuatorial contribuye con una energía favorable de 7.4 kcal/mol, mientras que con GalNAc, el OH axial contribuye con una energía de alrededor de 5.4 kcal/mol.

Tabla 4.4. Valores medios para la energía de interacción calculada con MM-GB(SA)

	CBS-B GlcNAc	CBS-B GalNAc
Ele	-75.3	-66.3
VDW	-15.2	-16.7
GAS	-90.5	-83.0
ΔG SV	61.2	61.9
GB$_{TOT}$	-29.1	-21.1

* Los valores de energía están expresados en kcal/mol.

4.3.3. Análisis de la estructura del solvatación sobre CBS-A de ABL

Como se ha mostrado en el capítulo previo, la determinación de la estructura del solvente en el CBS puede funcionar como un buen predictor de las características del complejo proteína-ligando.[5,6] El análisis esta basado en la definición y la caracterización estructural y termodinámica de los sitios de hidratación (WS) en la superficie del sitio de unión de carbohidratos (CBS) con mayor probabilidad de retener aguas. Dada la sutileza de la especificidad epimérica del CBS-A y CBS-B de ABL, proponer una explicación basada en análisis de las interacciones directas entre el CBS y los carbohidratos involucrados en base a las estructuras resueltas por difracción de rayos-X no resulta suficiente y por lo tanto es necesario evaluar el rol de las moléculas de aguas en la selectividad entre epímeros, debido principalmente a que los CBS están muy expuestos al solvente acuoso y los ligandos tienen un alto carácter hidrofílico.

Para analizar la estructura del solvente se han definido los posibles WS para ambos CBS y se han calculado las propiedades termodinámicas y estructurales de los mismos como fue descripto en el capítulo 3. De este modo, se han identificado 5 WS para el CBS-A y 4 WS para el CBS-B. Comenzamos el análisis inspeccionando la estructura del solvente del CBS-A. Como es posible observar en la Figura 4.6 A, en la cual se muestra el gráfico de la función de distribución radial (g(r)) de los átomos de oxígeno pertenecientes a las moléculas de agua con respecto a la Gly49-O, mientras que la Figura 4.6 B muestra el correspondiente gráfico bidimensional de distribución angular para las moléculas de agua en la primera capa de solvatación de Gly49-O y la Figura 4.6 C, los WS identificados en el CBS en torno a la región de la proteína que participa en la discriminación entre epímeros. Estos resultados muestran que Gly49-O tiene una primera capa de solvatación bien definida y el análisis de la estructura de solvatación en torno al mismo permite definir 3 WS bien establecidos. Cabe señalar que a pesar de que el WS2 en el gráfico bidimensional aparenta estar pobremente definido y localizado entre el WS1 y el WS3, el mismo es sumamente relevante, como se verá más adelante, para establecer la especificidad (los resultados de los parámetros estructurales y termodinámicos para los correspondientes WS se muestran en la Tabla 4.5)

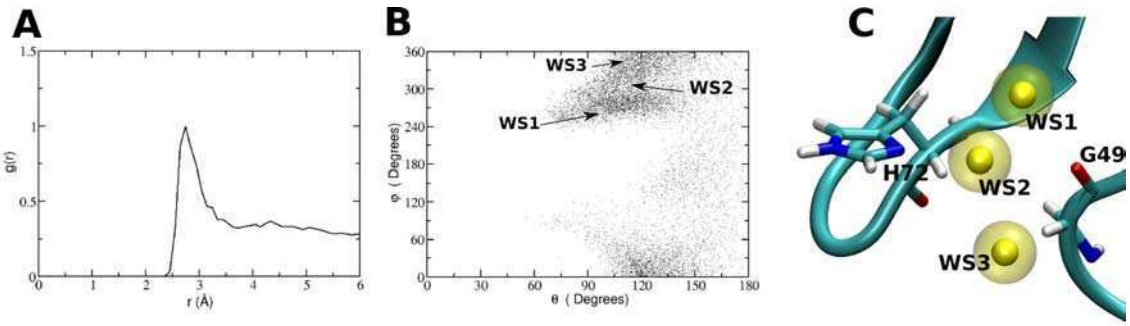

Figura 4.6. A) Gráfico de g(r) vs. la distancia entre el Gly49O y los átomos de oxígeno del solvente. **B)** Gráfico de distribución bidimensional para los átomos de oxígeno del solvente en que se encuentran en la primera capa de solvatación de Gly49O. **C)** WS sobre el CBS-A en torno a la región que define la especificad epimérica.

Tabla 4.5. Valores medios de energía de interacción, WFP y R_{min} para los WSs previamente definidos

Referencia	**WS**	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	**WFP**	**GalNAc**	**—**	**GlcNAc**	**—**
CBS-A	-	-	-	-	-	-	R_{min}	**átomo**	R_{min}	**átomo**
Gly 49 O	WS1	-17.0	-3.3	-20.3	-2.9	8.2	—-	—-	—-	—
Gly 49 O	WS2	-15.8	-4.8	-20.6	-3.2	4.5	2.4	O4	>5.0	O4
Gly 49 O	WS3	-15.7	-4.5	-20.2	-2.8	5.5	1.9	O4	3.2	O4
-	-	-	-	-	-	-	1.3	O3	1.9	O3
Asn 73 N	WS4	-7.1	-10.4	-17.5	-0.1	5.5	1.4	N7	1.4	N7
Ser 48 OG	WS5	-6.4	-13.0	-19.4	-2.0	5.3	0.7	O7	0.7	N7

* Los valores de energía y R_{min} están expresados en kcal/mol y Å, respectivamente.

Los resultados muestran que a pesar de que el WS1 se encuentra lejos de algún átomo del ligando en el complejo, los WS2 y WS3 se encuentran cercanos a la posición que ocupa el O4 axial. El WS3 es cercano al O4 axial en el complejo formado con GalNAc (el ligando natural), en donde las aguas que ocupan dicha posición forman el mismo patrón de puentes de hidrógeno y son desplazadas cuando GalNAc se une. Para el ligando invertido, GlcNAc, la posición del O4 ecuatorial se encuentra alejada de la posición predicha por el WS3 (y también del WS2), sugiriendo que el análisis de la posición de los WS refleja la especificad entre epímeros. Los WS4 y WS5 están muy cercanos entre sí, y posiblemente reemplazados por los átomos N7 y O7 del grupo N-Acetilo de ambos ligandos. Esto esta de acuerdo con el hecho de que el grupo N-Acetilo es el principal punto de anclaje que suelen tener los carbohidratos cuando actúan como ligandos y con los resultados del capítulo 3 que muestran que los WSs son capaces de predecir la posición de los grupos hidrofílicos.[5–7]

Para poder vincular la relación entre la posición de las moléculas de agua en la proteína libre y la movilidad de los grupos -OH del azúcar en el complejo, se han computado los funciones de distribución radial y angular para el OH4 del ligando, utilizando como referencia el átomo de oxígeno carbonílico de Gly49 (Gly490) con los dos ligandos (GalNAc y GlcNAc) y con la histidina protonada en el Nϵ (vale aclarar que para GlcNAc solo la primera parte de la dinámica fue utilizada dado que luego el ligando se suelta de la proteína -ver figura 4.4 A). Estos resultados se pueden observar en la figuras 4.7 A y B en donde se muestran los gráficos de distribución radial para el O4 de GalNAc y GlcNAc respectivamente tomando como referencia el oxígeno carbonílico de Gly49. Por otro lado, las figuras 4.8 A y B muestran las distribuciones angulares de los OH4 para cada ligando respectivamente.

Los datos muestran que para GalNAc la distribución radial presenta un primer pico bien definido a 3 Å de Gly49O, que es la misma distancia a la que se encuentra el pico de la primera esfera de solvatación en la proteína libre para Gly49O (ver figura 4.6 A). Por otro lado, para GlcNAc, la función de distribución radial para el O4 muestra un pico de baja probabilidad a 5 Å, mostrando claramente que el O4 ecuatorial no puede estar lo suficientemente cerca del O carbonílico de Gly49 para poder establecer una interacción estable. Una comparación adicional a partir de los gráficos de distribución angular (ver figuras 4.8 A y B) muestran claramente que el O4 axial de GalNAc solapa muy bien con la posición de las moléculas de agua dentro del WS3, pero el O4 ecuatorial de GlcNAc no tanto, reflejando nuevamente el rol predictivo de la interacción de las moléculas de agua en la proteína libre con lo consecuente formación del complejo. Como bien se ha mencionado anteriormente, aunque el WS2 no es desplazado por ninguno de los dos ligandos, juega un papel relevante en la especificidad. El análisis de la simulaciones de los complejos unidos a GalNAc y GlcNAc, revela la presencia de una molécula de agua formando una fuerte interacción entre la proteína y el ligando, actuando la misma como agua puente,

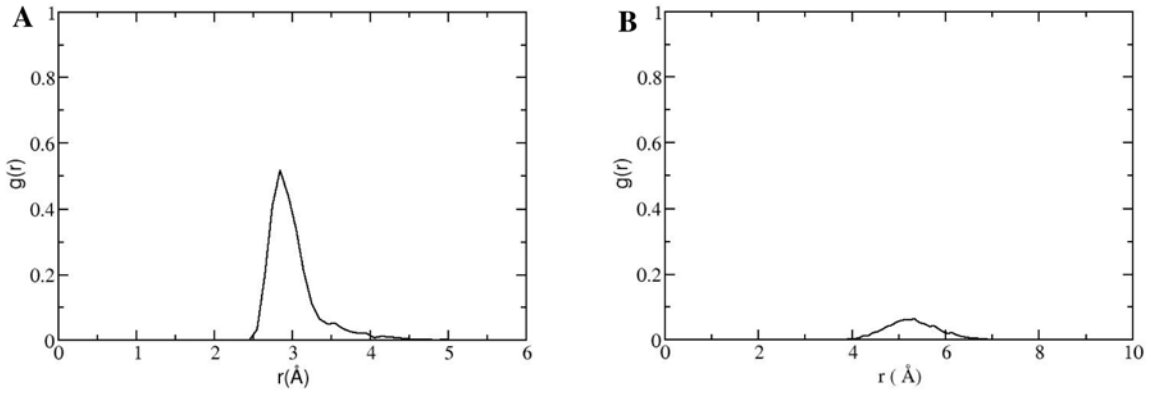

Figura 4.7. A) Gráfico de g(r) vs. la distancia entre el Gly49O y el OH4 GalNAc en el CBS-A. **B)** Gráfico de g(r) vs. la distancia entre el Gly49O y el OH4 GlcNAc en el CBS-A.

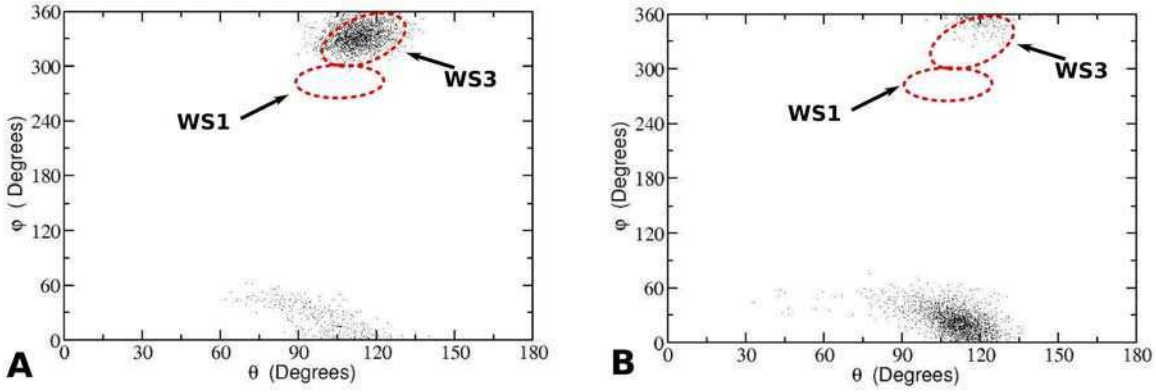

Figura 4.8. A) Gráfico de distribución angular bidimensional del oxígeno axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Gly49 O. **B)** Gráfico Bidimensional del oxígeno ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Gly49 O. Las regiones delimitadas por las líneas discontinuas rojas representan la posición de los WS, tal cómo se muestra en la figura 4.6 B.

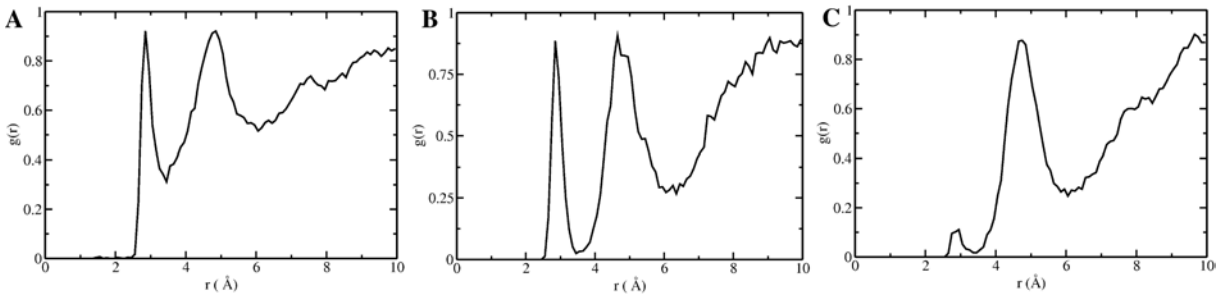

Figura 4.9. Gráfico de g(r) vs. la distancia de los átomos de oxígeno de las moléculas de agua del solvente con respecto al Nδ de la HIS72. **A)** Para el CBS-A sin ligando de la proteína ABL. **B)** Para el CBS-A unido a su ligando natural, GalNAc. C) Ídem B, pero en el complejo con GlcNAc unido al CBS-A

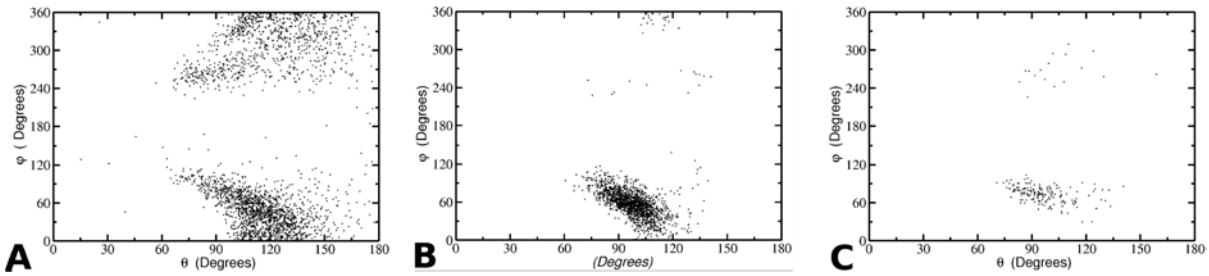

Figura 4.10. Gráfico de distribución angular bidimensional del oxígeno de las moléculas de agua del solvente con respecto al Nδ de la HIS72. **A)** Primera capa de solvatación en la proteína libre. **B)** Primera capa de solvatación en el complejo con β-GalNAc (es el observado experimentalmente). C) Ídem B, pero en el complejo con β-GlcNAc

coincidiendo perfectamente con la posición del WS2 y su interacción puente de hidrógeno con el Nδ de HIE72 y el O4 axial del ligando natural. Para analizar desde un punto de vista comparativo la estructura de este WS en el modelo libre (no unido a ligando), unido a GalNAc (ligando natural) y GlcNAc (ligando invertido) se calcularon las funciones de distribución radial de esta moléculas de agua para cada sistema respectivamente. Los resultados muestran claramente un pico bien definido con un alto valor de probabilidad en dicha zona para la proteína libre, el cual se observa más localizado (mas angosto y mayor valor de g(r)) cuando GalNAc (ligando natural) esta unido y prácticamente desaparece cuando GlcNAc (ligando invertido) se encuentra unido (ver figura 4.9 A, 4.9 B y 4.9 C respectivamente). Por otro lado, las figuras 4.10 A, 4.10 B y 4.10 C muestran la distribución angular de los átomos de oxígeno de las moléculas de agua, el OH4 axial y ecuatorial de GlcNAc y GalNAc con respecto al Nδ de HIE72, respectivamente. De las mismas se puede inferir que las moléculas de agua en el WS2 presentan menor movilidad en el complejo ABL-GlcNAc, respecto de ABL libre, mientras que en el caso del complejo invertido ABL-GalNAc, la presencia de la molécula de agua en el WS2 se observa en muy pocas fotos de

Tabla 4.6. Valores medios de la energía de interacción calculada con MM-GB(SA) considerando la contribución específica de la molécula de agua en el WS2

	Con H_2O	**Sin H_2O**	**ΔE**
Ele	-25.5	-19.1	-6.5
vdW	-17.8	-18.3	0.5
GAS	-43.3	-37.4	-5.9
ΔG SV	29.7	26.9	2.8
GB_{TOT}	-13.0	-10.5	-2.5

* Los valores de energía están expresados en kcal/mol.

la dinámica molecular. Adicionalmente, de manera coincidente a como se observa en la figura 4.10 B, el incremento en la estabilidad de las moléculas de agua cuando GalNAc esta unido, se evidencia a partir del cálculo de WFP, que se incrementa alrededor de 3 veces a un valor final de 12.0 con respecto al valor que toma en la proteína libre. Por último, con el propósito de determinar la contribución energética de la molécula de agua puente a la energía libre de unión global , se cálculo la energía de interacción entre GalNAc con la aproximación del modelo generalizado de Born (GB), pero incluyendo esta molécula de agua como parte del sistema analizado. Los resultados, que se pueden observar en la Tabla 4.6, muestran que la molécula de agua puente contribuye a aumentar la interacción electrostática en 6.5 kcal/mol y el valor total de la energía de interacción (incluyendo la parte de solvatación) en 2.5 kcal/mol, incrementando aún más las diferencias observadas en la Tabla 4.2.

4.3.4. Análisis de la estructura del solvatación sobre CBS-B de ABL

De manera similar a como se hizo en la sección anterior, es posible analizar la estructura de solvatación en el CBS-B utilizando como referencia el grupo carboxilato de Asp79, el cual interactúa de manera directa con el O4 de GlcNAc. La figura 4.11 A, muestra la función de distribución radial para los oxígenos de las moléculas de agua tomando como referencia el átomo de carbono (CG) del grupo carboxilato perteneciente a la cadena lateral de Asp79 y la figura 4.10 B, el correspondiente gráfico de distribución angular con la posición de los 3 WS identificados. Los parámetros termodinámicos y estructurales para los WSs correspondientes se muestran en la tabla 4.7.

Los resultados muestran que el WS1 tiene un valor de R_{min} de 1.0 Å respecto del OH4 del GlcNAc (el ligando con el que se cristalizo el complejo), con lo cual la molécula de agua que aloja este WS es expulsada hacia el seno del solvente cuando se forma el complejo, resultando ser un buen predictor epimérico, dado que además su posición no coincide perfectamente con la del OH4 cuando en su lugar se une el ligando invertido, GalNAc

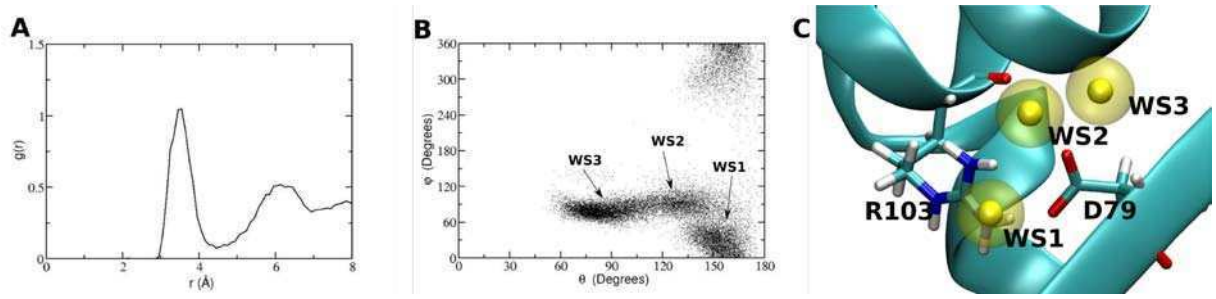

Figura 4.11. A) Gráfico de g(r) vs. la distancia entre el Asp79 CG y los átomos de O de las moléculas de agua del solvente. **B)** Gráfico de distribución angular bidimensional para los átomos de O del solvente en que se encuentran en la primera capa de solvatación de Asp79 CG. **C)** WS sobre el CBS-B en torno a la región que define la especificidad epimérica.

Tabla 4.7. Valores de energía de interacción, WFP y R_{min} para los WSs previamente definidos

Referencia	**WS**	$< E_p >$	$< E_w >$	$< E_T >$	$< \Delta E >$	**WFP**	**GalNAc**	**—**	**GlcNAc**	**—**
CBS-B	-	-	-	-	-	-	R_{min}	**átomo**	R_{min}	**átomo**
Asp 79 CG	WS1	-19.4	-3.4	-22.8	-5.4	11.0	1.0	O4	1.4	C4
Asp 79 CG	WS2	-14.0	-4.4	-18.4	-1.0	4.8	1.5	O4	2.8	C4
Asp 79 CG	WS3	-12.4	-2.1	-14.5	2.9	—	—	O4	4.9	O4
Thr 82 N	WS4	-6.9	-11.6	-18.5	-1.1	11.0	0.2	O7	0.7	O7

* Los valores de energía y R_{min} están expresados en kcal/mol y Å, respectivamente.

(R_{min} al OH4 mayor a 3 Å). Luego, el WS4 es perfectamente reemplazado por el O7 por ambos ligandos (GlcNAc y GalNAc) confirmando nuevamente el poder predictivo de los WS. Nuevamente, para comparar la estructura del solvente en el sistema libre con la estructura del ligando unido, se calcularon las funciones de distribución radial para el OH4 con los complejos formados por ambos azúcares. Los resultados, que se muestran en la figura 4.12 A y B, muestran que la posición del pico del OH4 para ambos ligandos es alrededor de 3.5 Å, que coincide con la distancia más probable de los átomos de oxígeno de las moléculas de agua respecto de la misma referencia (ver figura 4.11 A) y esto es consistente con el hecho de que Asp79 es capaz de interactuar con ambos epímeros (ver tabla 4.3). Por otro lado, el gráfico bidimensional mostrado en la figura 4.13, muestra diferencias significativas entre ambos complejos. Mientras que la posición media angular del O4 del ligando natural GlcNAc se encuentra entre la posición del WS1 y WS2 (ver figura 4.11 B), la posición del GlcNAc (ligando invertido) está solo sutilmente superpuesta con el borde del WS1 (ver figura 4.11 B y 4.13 B). En resumen, el análisis de la estructura de solvatación del receptor libre, muestra que es una herramienta útil para explicar el detalle de la selectividad epimérica en este caso, revelando una sutil preferencia para GlcNAc sobre GalNAc.

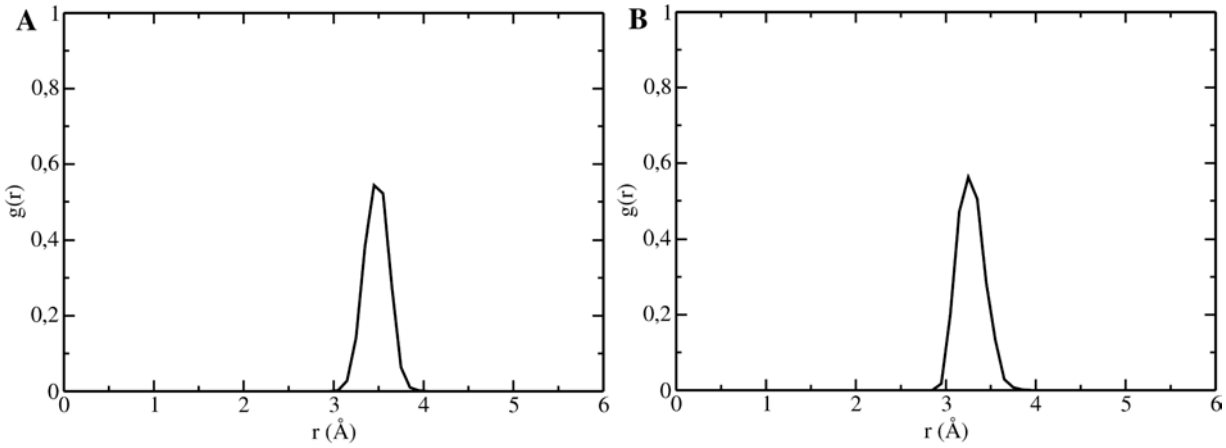

Figura 4.12. A) Gráfico de distribución radial, g(r) del O4 ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Asp 79 CG. **B)** Gráfico de distribución radial, g(r), del O4 axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Asp79 CG.

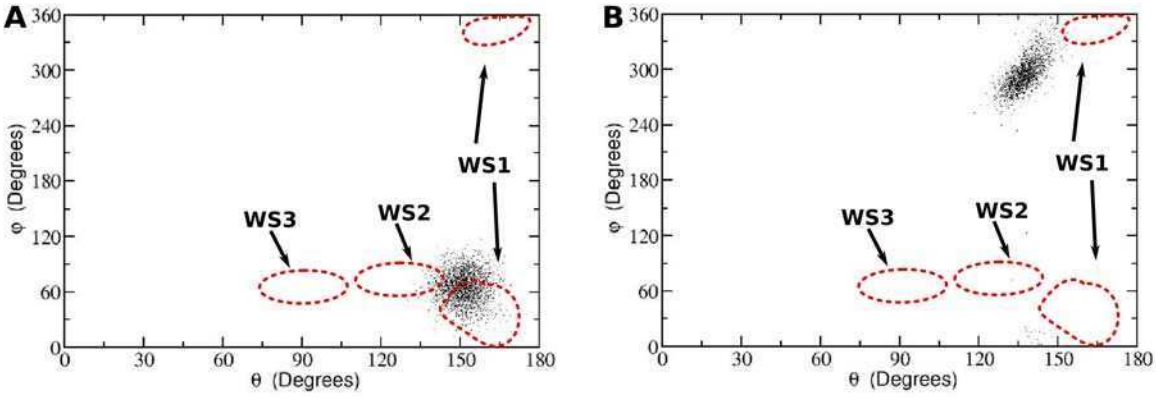

Figura 4.13. A) Gráfico de distribución angular bidimensional del O4 ecuatorial de β-N-Acetil-D-Glucosamina (GlcNAc) con respecto al Asp 79 CG. **B)** Gráfico de distribución angular bidimensional del O4 axial de β-N-Acetil-D-Galactosamina (GalNAc) con respecto al Asp79 CG. Las regiones delimitadas por las líneas discontinuas rojas representan la posición de los WS, tal cómo se muestra en la figura 4.11 B.

4.4. Discusión

Los dos CBS de ABL ofrecen la posibilidad única de estudiar como cada uno de ellos es capaz de distinguir dos monosacáridos muy similares entre sí, que difieren únicamente en la configuración del grupo hidroxilo en posición 4 (ver figura 4.1). Utilizando simulaciones de dinámica molecular de los complejos formados por los ligandos naturales e invertidos (en donde el ligando fue cambiado *In-Silico* a su epímero correspondiente), en combinación con el análisis de la estructura del solvente en la proteína libre, se han podido dilucidar las razones estructurales que definen la especificad epimérica y se ha podido medir de manera aproximada, aunque cuantitativa, el grado de selectividad epimérica que cada sitio presenta. Los resultados de las simulaciones complementan la evidencia experimental provista acerca de la especificidad epimérica que muestra ABL[1] permitiendo un análisis detallado de:

(a) Cada interacción en particular y su contribución relativa a la energía de unión global, que terminarán determinando la especificad epimérica.

(b) Una predicción semi-cuantitativa del grado de selectividad epimérica de cada sitio.

(c) Un análisis de la estructura de solvatación de la proteína libre, mostrando como la misma está relacionada con el complejo proteína-carbohidrato y podría, por consiguiente, ser utilizado como un predictor de especificidad.

En el contexto de la unión proteína-carbohidrato, el sitio de unión CBS-A de ABL une preferencialmente GalNAc sobre GlcNAc, de hecho los resultado de dinámica molecular muestran que dicho sitio es incapaz de unir GlcNAc. Las interacciones relevantes para mantener al carbohidrato en su lugar son las formadas entre el oxígeno carbonílico del grupo N-Acetilo del ligando con Ser48 y Asn73 de la proteína, más una interacción específica entre el OH4 la cual forma un puente de hidrógeno con el grupo carbonilo de Gly49 y una molécula de agua puente con el Nδ de la His72. Está molécula de agua puente es crucial en el estado tautomérico de la His72, HIE, la cual se mantiene unida a la proteína libre cuando la histidina tiene este estado de protonación y no en el HID (del mismo modo, como se observa en las figuras 4.9 y 4.10, esta molécula de agua solo está presente cuando el ligando natural, β-GalNAc, está unido al CBS-A). **Por otro lado, dado que solo se establecen entre 3 y 4 interacciones puentes de hidrógeno entre la proteína y el carbohidrato, cambiar la posición del OH4 como sucede cuando se une GlcNAc, redunda en la pérdida de 1/2 a 2/3 de las interacciones lo cual no hace viable la unión del ligando invertido, determinando de esta manera la especificidad epimérica**. Esto es consistente con la evidencia experimental que muestra que solamente existe densidad electrónica en el sitio

de unión cuando GalNAc esta presente, mientras que esto no sucede si se utiliza Galactosa o GlcNAc.[1] Es importante resaltar que la pérdida de unión de galactosa pone de relieve la importancia de las interacciones puentes de hidrógeno formadas por el grupo N-Acetilo.

En comparación con el CBS-A, la preferencia epimérica del CBS-B parece mucho menos pronunciada. Ambos complejos, el que tiene al ligando natural, GlcNAc, y el ligando invertido mutado *In-Silico*, GalNAc, se observan estables en el lapso de dinámica molecular simulado. En conjunto, el análisis de puentes de hidrógeno muestra para el ligando natural se establecen 4 interacciones estables, que están presentes entre el OH3, OH4, OH6 y carbonilo del grupo N-Acetilo. El hecho de cambiar GlcNAc por GalNAc da como resultados un corrimiento en la posición relativa del carbohidrato, redundando en la pérdida de la interacción del grupo carboxilato con el OH4, lo que concomitantemente debilita las otras interacciones. **La diferencia en la energía libre de unión predicha entre el complejo natural, GlcNAc, y el invertido, GalNAc, es de 8 kcal/mol correspondiendo a alrededor de 1/4 de la energía de interacción del ligando natural. Estos resultados sugieren que el CBS-B une GlcNAc pero también GalNAc con menor afinidad**. Sin embargo, los datos adquiridos a partir de los experimentos de difracción de rayos-X,[1] no muestran densidad electrónica en el CBS-B cuando los cristales son embebidos con GalNAc. Esto puede ser parcialmente explicado por el hecho que en estos cristales de ABL, solamente el CBS-A esta libremente accesible, mientras que el CBS-B en cambio, se muestra parcialmente bloqueado por el empaquetamiento del cristal, como se evidencia a partir de la ruptura del cristales cuando se embeben en GlcNAc. De hecho, obtener los cristales de los complejos ABL-GlcNAc lleva más tiempo y tienen una forma cristalina diferente en la cual el CBS-B no se encuentra bloqueado. Para testear si el CBS-B es capaz de unir GalNAc y evaluar más detalladamente su capacidad de reconocer epímeros, se requieren futuros experimentos. Una hipótesis que es posible plantear en base a los resultados presentados en este capítulo, consiste en mutar uno y/o dos residuos clave del CBS-A (His72 y/o Ser48) para inhibir la unión de GalNAc al CBS-A y de esa manera evaluar la unión de GalNAc a estas mutantes.

Si comparamos ahora los dos CBS entre sí, se observa que la principal diferencia entre el CBS-A y el CBS-B es la presencia del grupo ácido del Asp79 en el CBS-B. Este grupo cargado negativamente hace posible la formación de dos puentes de hidrógeno muy estables, los cuales resultan en una alta energía de interacción electrostática. Este alto valor de energía electrostática, a pesar de ser parcialmente compensada por la energía de desolvatación, redunda en una mayor energía de interacción predicha en el CBS-B comparada con el CBS-A. Para confirmar esto, se requiere hacer medidas experimentales de las constantes de disociación/asociación o bien medidas de la constante de equilibrio. Las contribuciones de vdW son similares y de alrededor de -15 kcal/mol en ambos casos (ver tabla 4.2 y 4.4). La diferencia en la especificidad epimérica es mayor para el CBS-A principalmente por dos

razones; la primera de ellas es que en el CBS-A el OH4 establece dos interacciones puente de hidrógeno con el ligando natural, las cuales se pierden cuando se utiliza su epímero. En el CBS-B, solamente una interacción puente de hidrógeno esta presente entre el OH4 y la proteína. En segundo lugar, el número total de puentes de hidrógeno es menor para el CBS-A y por consiguiente, perder una interacción tiene un efecto relativo mayor que para el otro CBS. Es interesante hacer notar que a pesar de que en muchas lectinas, como Gal-1,[8] las caras no polares de los carbohidratos interactúan con los residuos aromáticos de la proteína, estableciendo un interacción hidrofóbica del tipo π-stacking, para esta lectina, ninguna de estas interacciones se encuentran en ninguno de los dos CBS. Como se ha descripto en la introducción y en el capítulo anterior, la estructura del solvente en la proteína libre emula algunas de las propiedades estructurales de los complejos proteína-carbohidrato, dado que aquellas regiones del espacio con alta probabilidad de alojar moléculas de agua en la superficie proteíca, tienden a ser ocupadas por los grupos hidroxilos de los carbohidratos.[5,6]

Por otro lado, la discriminación entre epímeros depende de sutiles diferencias estructurales y por consiguiente, representa un caso testigo interesante para examinar cuán bien la estructura del solvente puede predecir la conformación correcta del complejo proteína-carbohidrato. Estos resultados muestran que el análisis de la estructura del solvente, a través de la aproximación de los WS, puede en cierta medida predecir o al menos explicar la preferencia entre epímeros. **Para el CBS-A los WS con un WFP aproximadamente mayor a 5 (WS3), están posicionados (como se evidencia a partir de los gráficos bidimensionales) en el mismo lugar que el OH4 de GalNAc pero no de GlcNAc. Además, las moléculas de agua en el WS2 actúan como puente entre el OH4 de GalNAc, contribuyen significativamente a mantener al ligando unido pero que se pierde cuando GlcNAc es forzado a unirse. Para el CBS-B, el análisis de la posición del WS2 muestra que puede ser desplazado por el OH4 ecuatorial de GlcNAc y esta lejos de GalNAc, y por consiguiente, un análisis cuidadoso de la estructura del solvente revela la especificidad epimérica**.

En resumen, estos resultados sostienen la hipótesis de que la estructura del solvente permite predecir ciertas características de los complejos proteína-carbohidrato e incluso ser capaces de determinar las posibles selectividades epiméricas. Sin embargo, al momento de evaluar la selectividad, es importante tener en cuenta tanto las interacciones que el ligando puede establecer con el CBS del receptor con los grupos funcionales del ligando, así como los cambios en la estructura de solvatación de ambas biomoléculas.

4.5. Conclusión

Desde un punto de vista general, los resultados mostrados resaltan la dificultad y la sutileza del proceso de reconocimiento de carbohidratos. Los mismos muestran que solamente para el CBS-A, donde 2 de las 4 interacciones puente de hidrógeno para el ligando natural se pierden, la proteína es capaz de discriminar eficientemente entre los dos epímeros unidos. Este no es el caso para el CBS-B donde solo alrededor de 1 puente de hidrógeno se pierde y tanto GlcNAc como GalNAc se mantienen unidos. Es interesante notar que, los datos de rayos-X (PDB ID 1Y2U) muestran que a pesar de que la inversión de ligandos en los CBSs no ocurre y que GlcNAc no se une al CBS-A, se observa que un disacárido conteniendo GlcNAc y Galβ1 (Lacto-N-Biosa) se une a un sitio adyacente y forma un complejo estable. El segundo GalNAc agrega entre 3 y 4 puentes de hidrógeno y por consiguiente, compensa las pérdidas de interacción descriptas previamente. Esto muestra que la selectividad epimérica puede ser solo alcanzada por monosacáridos y requiere, como es de esperar, un ajuste fino de las interacciones entre la proteína y el HO4 en combinación con una unión no demasiado fuerte, dado que sitios que son capaces de establecer numerosas interacciones con el ligando, como en el CBS-B, o para el caso de unión de disacáridos en el CBS-A, resulta en un decremento de la selectividad.

Aunque es necesario establecer un análisis exhaustivo sobre todos los casos posibles para llegar a una conclusión general, los resultados presentados en este capítulo sugieren que la selectividad de carbohidratos puede ser difícil de alcanzar, sugiriendo una elección cuidadosa de la configuración de los sustituyentes en los ligandos glicomiméticos no naturales.

4.6. Metodología específica del capítulo

Las estructura inicial correspondiente al complejo simulado que tiene los dos ligandos unidos fue generada partiendo de la estructura cristalográfica de GalNAc unido al sitio A de ABL y GlcNAc unido al sitio B de ABL (PDBID 1Y2X) y a partir de ahora referido como ABL-GalNAc-A y ABL-GlcNAc-B. Además, ambas estructuras epiméricas correspondientes a GalNAc unido al sitio B (ABL-GalNAc-B) y GlcNAc unido al sitio A (ABL-GlcNAc-A) fueron construida *In-Silico* cambiando la posición correspondiente del grupo OH y H que estuvieran unidos al C4 del monosacárido, dando como resultado el complejo di-ligando invertido. Como se evidencia experimentalmente,[1] ambos CBS son independientes y por lo tanto, la unión de un carbohidrato a uno de ellos, no afecta las propiedades de unión a ligandos del otro. De este modo, solo fue necesario realizar dos simulaciones de dinámica molecular, una para los ligandos en su estado natural y otra para los ligando invertidos, con el fin de

reducir el costo computacional.

Dado que el residuo 72 en el CBS-A es una histidina, se tuvo en cuenta el estado de protonación en sus distintos estados tautoméricos, HIE (protonada en el Nϵ) y HID (protonada en el Nδ) y de este modo, los dos posibles estados fueron generados (ABL-GalNAc-A-HIE y ABL-GalNAc-HID, respectivamente). Finalmente, la proteína ABL en su estado libre de ligandos fue generada a partir de remover los carbohidratos de la estructura cristalograficas utilizada para construir los complejos y se simulo la misma con la HIS 72 en su estado tautomerico HIE. Cada uno de estos 5 sistemas fue sometido al mismo protocolo de equilibración y de producción de dinámica molecular que se describe en la sección Metodología específica del capítulo 3. Para caracterizar las interacciones proteína-ligando y/o agua, se realizo un análisis de puentes de hidrógeno entre los átomos del ligando y la proteína como ser la distancia media y el ángulo medio entre el átomo donor y aceptor de puentes de hidrógeno y se consideró que una interacción de este tipo esta presente cuando la distancia donor-aceptor es menor a 3 Å y el ángulo donor-hidrógeno-aceptor es menor a 120^0.

Referencias

[1] M.E. Carrizo, S. Capaldi, M. Perduca, F.J. Irazoqui, G.A. Nores, and H.L. Monaco. The antineoplastic lectin of the common edible mushroom (agaricus bisporus) has two binding sites, each specific for a different configuration at a single epimeric hydroxyl. *Journal of Biological Chemistry*, 280(11):10614–10623, 2005. cited By (since 1996)42.

[2] G.F. Springer. Immunoreactive t and tn epitopes in cancer diagnosis, prognosis, and immunotherapy. *Journal of Molecular Medicine*, 75(8):594–602, 1997. cited By (since 1996)208.

[3] L.-G. Yu, N. Andrews, M. Weldon, O.V. Gerasimenko, B.J. Campbell, R. Singh, I. Grierson, O.H. Petersen, and J.M. Rhodes. An n-terminal truncated form of orp150 is a cytoplasmic ligand for the anti-proliferative mushroom agaricus bisporus lectin and is required for nuclear localization sequence-dependent nuclear protein import. *Journal of Biological Chemistry*, 277(27):24538–24545, 2002. cited By (since 1996)20.

[4] L. Yu, D.G. Fernig, J.A. Smith, J.D. Milton, and J.M. Rhodes. Reversible inhibition of proliferation of epithelial cell lines by agaricus bisporus (edible mushroom) lectin. *Cancer Research*, 53(18):4627–4632, 1993. cited By (since 1996)116.

[5] S.D. Lella, M.A. Martí, R.M.S. Álvarez, D.A. Estrin, and J.C. Díaz Ricci. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *Journal of Physical Chemistry B*, 111(25):7360–7366, 2007. cited By (since 1996)15.

[6] D.F. Gauto, S. Di Lella, C.M.A. Guardia, D.A. Estrin, and M.A. Martí. Carbohydrate-binding proteins: Dissecting ligand structures through solvent environment occupancy. *Journal of Physical Chemistry B*, 113(25):8717–8724, 2009. cited By (since 1996)13.

[7] D.D. Leonidas, B.M. Swamy, G.N. Hatzopoulos, S.J. Gonchigar, V.B. Chachadi, S.R. Inamdar, S.E. Zographos, and N.G. Oikonomakos. Structural basis for the carbohydrate recognition of the sclerotium rolfsii lectin. *Journal of Molecular Biology*, 368(4):1145–1161, 2007. cited By (since 1996)20.

[8] S. Di Lella, L. Ma, J.C. Díaz Ricci, G.A. Rabinovich, S.A. Asher, and R.M.S. Álvarez. Critical role of the solvent environment in galectin-1 binding to the disaccharide lactose. *Biochemistry*, 48(4):786–791, 2009. cited By (since 1996)11.

Capítulo 5

Mejora de los métodos de predicción de la estructura lectina-carbohidrato

5.1. Introducción

La potencialidad y la utilidad de los métodos para predecir la estructura de los complejos proteína-carbohidrato a partir de la conformación de ambas biomoléculas por separado, conocidos como métodos de docking, reside en su capacidad de predecir correctamente la estructura del complejo bimolecular y por consiguiente, las interacciones entre ambas subunidades, tomando como punto de partida la estructura de la proteína y del ligando en su estado no unido. A pesar de que actualmente, existen muchos programas de docking disponibles,[1–3] dada la gran cantidad de aproximaciones empleadas por los mismos, los resultados alcanzados no son siempre satisfactorios,[4–11] y aunque en muchos trabajos se compara su eficacia, aún no queda claro cuál es la mejor elección.

En particular, para cálculos de docking donde el ligando es un azúcar, el trabajo de Mishra et al. (2012) muestra que Autodock3 funciona mejor que Autodock4, Vina y DOCK, pero que da un gran número de falsos positivos como posibles resultados.[4,12–14] Por otro lado, se muestra que Glide[15] y Autodock4[16] tienen una mejor eficacia en predecir el complejo correcto, pero los resultados son fuertemente dependientes del par receptor-ligando que se está evaluando. De este modo, uno de los programas más ampliamente utilizado que además es libre bajo GNU (General Public Licence) es Autodock4, el cual utiliza un esquema de docking para la búsqueda de la estructura del complejo proteína-ligando a través un algoritmo genético[16] para explorar las posibles conformaciones que el ligando puede adoptar, en combinación la evaluación de la energía libre de unión para cada posible conformación del mismo en el complejo, la cual es calculada a partir de una función empírica cuyos parámetros fueron previamente ajustado sobre un amplio conjunto de valores de energía libre de unión de complejos proteína- ligando determinados experimentalmente.[1,17]

En el capítulo 3, en base a los resultado provistos por la Dinámica Molecular de las proteínas libres de ligando en solvente explícito, en combinación con estrategias de análisis

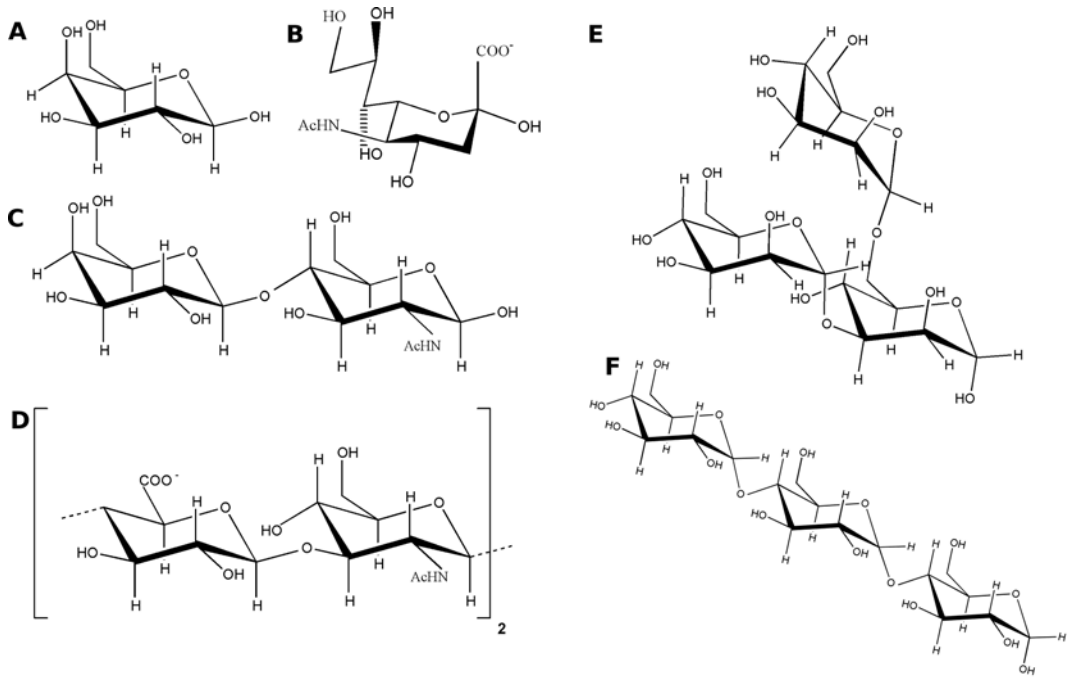

Figura 5.1. Esquema de los carbohidratos utilizados en el presente capítulo. Los pares proteína-ligando son: A) CBM32-Gal; B) CBM40-ácido siálico; C) Gal3-LAcNAc; E) ConA-trimanosido; D) CD44-HA_4; F) SPD-Maltotriosa

basadas en la termodinámica estadística, se ha mostrado cómo dilucidar la estructura del solvente en el CBS . Esto ha permitido calcular la energía de interacción de las moléculas de agua en el WS con la proteína y con su entorno acuoso, la probabilidad relativa de encontrar una molécula de agua en el WS con respecto al seno del solvente y el radio R_{90} (tomado desde el centro del WS) tal que comprenda al 90 % de las moléculas de agua que visitan el sitio en la dinámica. Finalmente se relacionaron estos parámetros, con la chance de que un grupo funcional del ligando desplace las moléculas de agua que ocupan los WS, a través del parámetro R_{min} concluyéndose que WFP y R_{90} son los mejores predictores para determinar la posición de los mismos. Estas observaciones han motivado el desarrollo del presente capítulo, en donde se mostrará cómo, utilizando la información provista desde la identificación y caracterización de la estructura de solvatación en los sitios de unión a carbohidratos (**CBS**) para un conjunto de proteínas que unen este tipo de ligandos, se puede modificar la función de energía de Autodock4[1,2,16,17] con el objetivo de mejorar la predicción *In-Silico* correspondiente a la estructura proteína-ligando. Finalmente, con el propósito de evaluar la eficiencia de la estrategia que se está presentando, se seleccionaron seis complejos proteína-carbohidrato cuya estructura cristalográfica es bien conocida, y cuyos ligandos comprenden un amplio rango de tamaños, desde los mono- hasta los tetrasacáridos (ver figura 5.1), comparándose en todos los casos con el método convencional (es decir, sin utilizar la información de los sitios de hidratación (WS)).

5.2. Ideas e hipótesis

Los métodos de docking *In-Silico* tienen una larga tradición en predecir la estructura de los complejos proteína-ligando, y permiten una rápida exploración de un número considerable de posibles estructuras. Sin embargo, determinar cuál es la correcta, no es siempre sencillo. En este capítulo, se modificarán las funciones de puntuación de Autodock4, un conocido y ampliamente utilizado programa de docking *In-Silico*, sobre la base de la información acerca de la estructura del solvente adyacente a la superficie proteica que se obtiene a partir de las simulaciones de dinámica molecular del receptor libre.

En base a los resultados de los capítulos previos y a la intima relación existente entre las superficies hidrofílicas del CBS de lectinas y los ligandos sacarídicos con el entorno acuoso, se espera que esta estrategia de docking *In-Silico* contribuya de manera significativa a determinar la estructura del complejo lectina-carbohidrato correcto con mayor exactitud y precisión que el método de docking convencional.

5.3. Resultados

Los resultados están organizados de la siguiente manera: Para cada caso puesto a prueba, se describe brevemente la relación entre la estructura de solvatación (WS) determinada (un resumen de los WS utilizados para construir las grillas modificadas se presentan en la tabla 5.1) para el CBS de cada proteína, con la estructura la posición del los átomos del ligando en la estructura del complejo. Luego se presentan los resultados obtenidos según el esquema de docking convencional (**CADM**, por sus siglas en inglés *"Conventional Autodock4 Docking Method"*) y con el esquema sesgado por los sitios de aguas (**WSBDM**, por sus siglas en inglés *"Waters Sites Bias Docking Method"*), siempre usando la estructura cristalográfica del complejo como referencia (en la sección "Protocolo y metódo de docking sesgado (WSBDM)" del capítulo de métodos se presenta una descripción detallada de la metodología empleada para este fin) . Los casos están organizados por tamaños crecientes de ligandos (desde mono hasta tetra sacáridos). Para algunos casos en particular se han variado parámetros como el número de WS utilizados para construir la grilla, la estructura del receptor y el tamaño del ligando, con el objetivo de encontrar los condiciones adecuadas en las que el WSBDM alcance su mayor carácter predictivo. En particular, para poder analizar visualmente los resultados de cada cálculo de docking se construyeron gráficos bidimensionales de Población vs. ΔG_b (en donde cada punto es un complejo propuesto por el programa), siendo estos dos, parámetros claves para poder discernir la conformación del complejo proteína-ligando que tenga mayores chances de coincidir con el complejo verdadero (es decir, aquel que se infiere a partir de resultados experimentales, como por

ejemplo, difracción de rayos-X). De este modo, el complejo correcto debiera ser aquel que según el cálculo de docking tenga el ΔG_u más negativo y el valor de Población más alto (o bien, una relación de compromiso entre ambos).

En este análisis, se espera observar en los gráficos bidimensionales construidos, que el complejo correcto se presente como un punto significativamente separado de los otros (los valores más negativos de ΔG_u y más grandes de población), con el fin de poder discriminarlo de los resultados falsos positivos. Por otro lado, para poder discernir de manera cuantitativa si los resultados del cálculo de docking coinciden con el complejo correcto, se calculó el valor de RMSD de los átomos de carbono del ligando sacarídico para los complejos predichos con respecto a la posición del carbohidrato que se desprende de la estructura cristalográfica. Finalmente, se incluirá una discusión global de los resultados presentados en este capítulo.

5.3.1. Docking de monosacáridos sobre los módulos CBM32 y CBM40 de la Sialidasa Bacteriana de *Clostridium Perfringens*

Como se ha descripto en el capítulo 3, se puede identificar en el CBS de CBM32, 4 WS, de los cuales tres (WS1, WS2 y WS4) son desplazados cuando la lactosa ocupa el sitio de unión (ver tabla 5.1 y figura 3.1 A y B para más detalles). Utilizando esta información de la estructura de solvatación, se ha re-dockeado la galactosa sobre la estructura cristalográfica del complejo CBM32-galactosa utilizando el método de docking convencional y el método de docking modificado (ver figura 5.2). Como puede observarse en la figura 5.2 A, en donde se gráfica la población vs. ΔG_b para los complejos predichos por el esquema de docking convencional, el complejo propuesto con mayor exactitud (es decir, aquel complejo en donde la posición del ligando tenga el menor valor de RMSD respecto de la estructura de referencia cristalográfica) se encuentra en una región del gráfico en donde es muy difícil distinguirlo del resto de los otros complejos. Otro resultado interesante, es que el punto de mayor valor de población (alrededor del 50 % de los complejos predichos son representados por este punto) posee un valor de RMSD de 3.5 Å. Sin embargo, un análisis más detallado de la estructura, ver figura 5.2 C, muestra que el complejo predicho por el programa de docking se encuentra rotado y desplazado de la estructura de referencia, lo cual nos muestra, que el método de docking convencional predice de manera incorrecta la estructura del complejo.

Los resultados que se han obtenido por el WSBDM, utilizando los 4 WS mencionados, se muestran en la figura 5.2 B. En este caso, solo dos grupos son encontrados, entre los cuales la diferencia de energía de unión no resulta significativa (la diferencia es de solo 0.2 Kcal/mol) y los valores de población son muy similares. Se observa que el complejo más cercano a la estructura cristalográfica (RMSD 1.1 Å), que se muestra en la figura 5.2 E, no

Tabla 5.1. Propiedades de los WS de los CBS para cada proteína utilizada

Proteína	**WS**	**WFP**	$\mathbf{R_{90}}$	$\mathbf{R_{min}}$
CBM32	1	18.9	1.6	1.5
	2	2.4	4.4	2.5
	3	6.2	2.1	0.9
	4	11.5	1.7	1.2
CBM40	1	17.8	1.8	0.8
	2	8.5	3.1	1.7
	3	13.6	1.7	0.7
	4	7.5	2.1	1.4
	5	6.5	3.5	1.3
	6	18.8	1.7	0.8
Galectina 3	1	7.0	2.2	1.6
	2	9.1	1.5	0.4
	3	4.1	3.3	1.8
	4	18.5	1.5	0.3
	5	7.0	2.4	1.3
	6	6.6	4.1	2.3
Concanavalina A	1	7.7	2.9	1.5
	2	8.8	2.8	1.2
	3	12.8	1.9	1.4
	4	1.1	1.4	1.5
	5	7.2	3.9	0.9
	6	9.3	2.2	1.5
	7	9.8	3.1	1.1
	8	20.3	1.6	0.7
	9	16.9	2.1	0.6
	10	3.4	2.5	2.9
	11	3.1	1.8	1.9
SP-D	1	3.7	0.8	0.4
	2	8.0	0.9	0.7
	3	2.4	0.6	2.1
	4	22.0	0.7	0.3
	5	24.1	0.6	1.8
CD44	1	16.0	0.8	0.7
	2	5.7	0.8	1.0
	3	5.4	0.6	0.5
	4	12.7	1.0	0.8
	5	2.7	0.4	2.1

es el de mayor valor de población, ni menor valor de ΔG_u. No obstante, el complejo que presenta el menor valor de energía muestra que la β-D-Galactosa se encuentra sutilmente desplazada y rotada alrededor de 60^0 respecto de la estructura de referencia, como puede observarse en la figura 5.2 D. A pesar de que para el método de docking modificado, la estructura correcta se ubica en segundo lugar en el ranking, funciona mejor que el método convencional.

Ahora pondremos nuestra atención en CBM40, la cual presenta 6 WS bien definidos en su sitio de reconocimiento de carbohidratos (ver tabla 5.1 y figura 3.2 A y B para más detalles). Tres de estos sitios tienen un alto valor de WFP (WS1, WS3 y WS6) y son desplazados por el O8, C6 y el grupo carboxilato respectivamente. La figura 5.3 A muestra el gráfico de Población vs. ΔG_b para el cálculo de docking del ácido siálico sobre el receptor de la estructura cristalina de CBM40 utilizando el método de docking convencional. Los resultados muestran que el metódo de docking convencional, CADM, rankea en primer lugar al complejo incorrecto ya que su valor de RMSD es de 4.7 Å respecto de la referencia, pero el segundo complejo en el ranking es el correcto (RMSD 1.1 Å). Por otro lado, con él método de docking modificado, WSBDM, (en donde se han utilizado los 6 WS para modificar la grilla), el cálculo muestra una mejor eficacia predictiva. El primer complejo en el ranking de energía y además con población más alta, es el complejo correcto, con un valor de RMSD de 0.9 Å. La estructura predicha muestra al esqueleto carbonado del monosacárido correctamente ubicado y orientado, con su principal cadena lateral levemente desplazada. No obstante, es importante resaltar que a pesar que su valor de población es similar al de los otros cluster que predicen de manera incorrecta el complejo cristalográfico, es posible elegirlo en base a que presenta el menor valor de energía de unión. En resumen, los resultados obtenidos para el docking de monosacáridos de los módulos CBM de la sialidasa de bacteria muestran una mejora significativa utilizando el WSBDM.

5.3.2. Galectina 3

Gal-3 muestra un CBS que por lo general, une a los disacaridos Lactosa o N-Acetil-Lactosamina, el cual presenta 6 WS (ver tabla 5.1 y figura 3.3 A y B para más detalles). Dos WS se encuentran claramente ubicados en el sitio de unión de Galactosa, otros tres WS en el sitio de unión de N-acetil-lactosamina y por último 1 WS se localiza entre los dos monosacáridos sobre la unión glicosídica. Los dos WS que muestran los más altos valores de WFP son claramente reemplazados cada uno de ellos por el O3 del grupo -OH de GlcNAc y el O6 de la Galactosa respectivamente. Los otros WS también se encuentran cerca de los grupos hidroxilo de ligando a excepción de WS6, que se encuentra cerca del CH_3 del grupo acetilo.

En un caso real es muy difícil conocer en que lugar se unirá cada monosacárido. Por

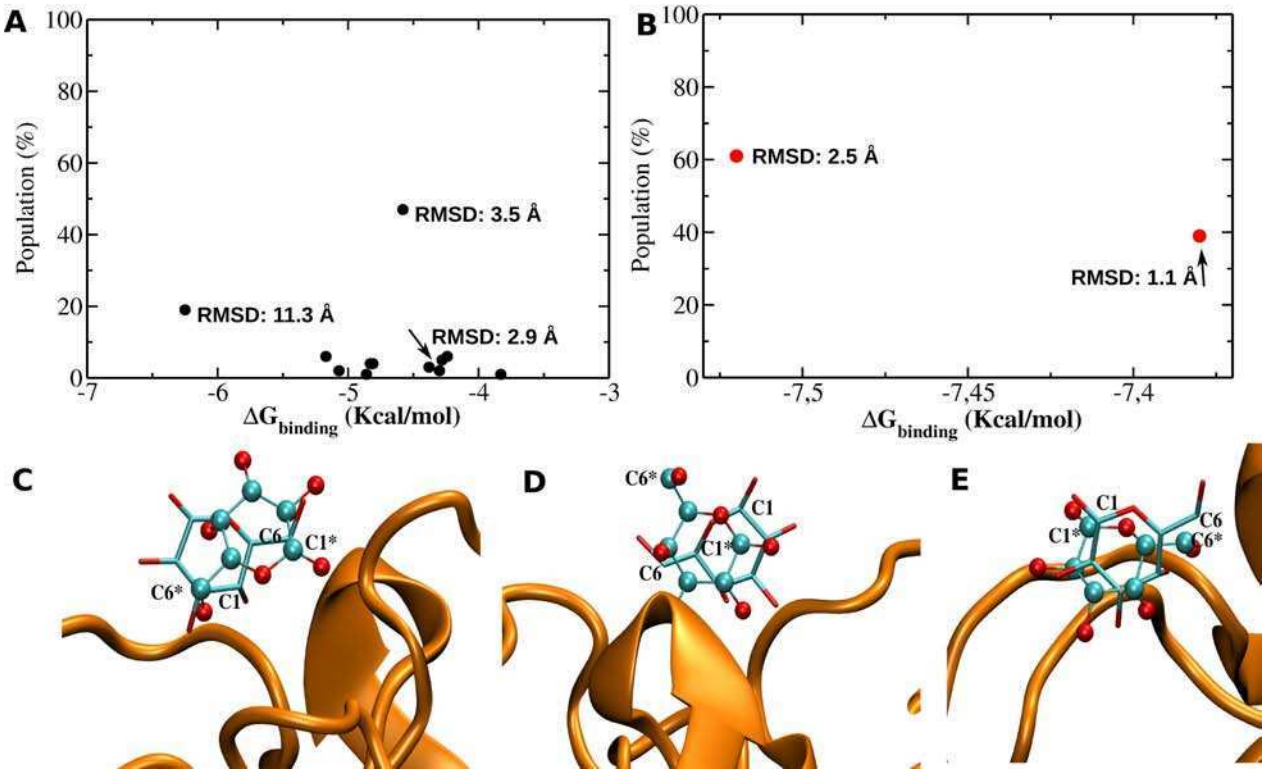

Figura 5.2. Resultado para el cálculo de docking de Gal sobre la estructura cristalográfica de CBM32. Se muestra la Población vs. ΔG_u utilizando **(A)** el método convencional de docking (CADM) y **(B)** el método modificado (WSBDM). Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono del ligando entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 2v72). En los paneles inferiores, se muestra la estructura predicha para el complejo CBM32-Gal superpuesta sobre la estructura de referencia. **(C)** La estructura correspondiente al complejo de población más alta con el CADM predicha con un valor de RMSD de 3.5 Å, **(D)** el complejo predicho de más baja energía con el WSBDM con un RMSD de 2.5 Å y **(E)** el segundo complejo predicho de más baja energía con el WSBDM con un RMSD de 1.1 Å. La posición de los ligandos en los complejos predichos por el programa de docking son mostrados en representación de "balls and sticks" mientras que la estructura de referencia se representa en "Licorice". Los átomos marcados con un asterisco (en las figuras C, D y E) corresponden a la estructura predicha, mientras que los que no están marcados con asterisco, a la estructura de referencia.

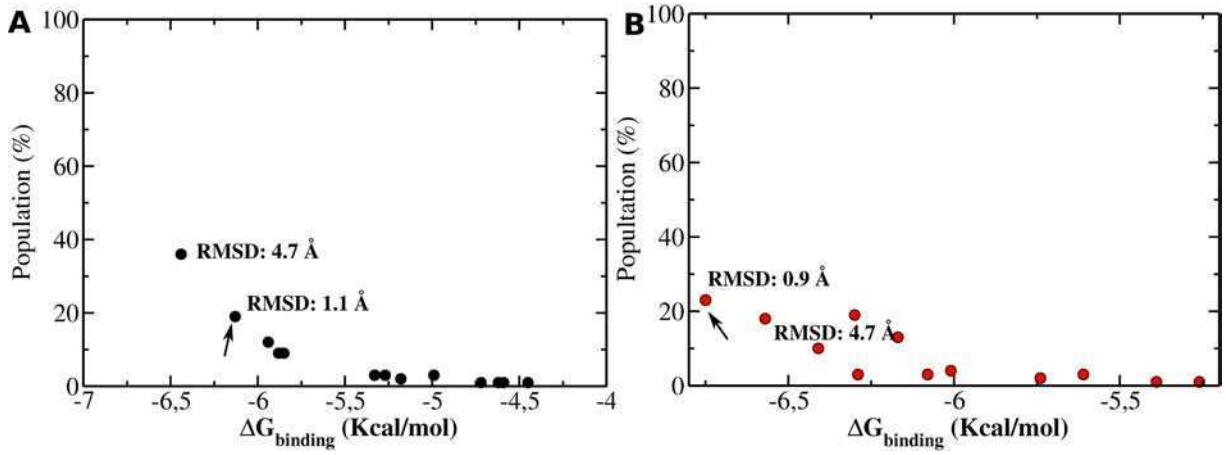

Figura 5.3. Resultado para el cálculo de docking del ácido siálico sobre la estructura cristalográfica de CBM40. Se muestra la población vs. ΔG_u utilizando **(A)** el método convencional de docking (CADM) y **(B)** el método modificado (WSBDM). Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 2v73)

consiguiente, aún para el caso de dockear un monosacárido, todo el CBS y sus WS asociados fueron analizados. Con esto en mente, se comparará si él método de docking convencional (CADM) y el método modificado (WSBDM) pueden determinar correctamente el complejo Gal3-disacárido. De este modo, se comenzará el análisis dockeando individualmente cada monosacárido (Gal y GlcNAc) sobre todo el sitio de reconocimiento de carbohidratos de Gal-3, luego se utilizará el disacárido completo y finalmente se evaluará la influencia en el número de WS sobre el eficacia del WSBDM.

5.3.3. Docking de Galactosa (Gal) sobre Gal-3

Los resultados para el docking de Gal sobre Gal-3, utilizando el CADM y WSBDM, sobre la estructura del cristal para construir la grilla que representa al receptor, son mostrados en la figura 5.4. Los resultados obtenidos para el CADM se muestran en la figura 5.4 A en donde se puede observar un complejo de alta población (46 %) y baja energía (señalado con una flecha), y con un valor de RMSD respecto de la referencia cristalográfica de 3.3 Å. Este es el complejo predicho que menor valor de RMSD presenta y muestra claramente que el carácter predictivo del método convencional no es satisfactorio. Por otro lado, si se utiliza el WSBDM, (ver figura 5.4 B), considerando los 6 WS encontrados en el CBS, los resultados muestran una mejor eficiencia para discernir el complejo correcto de los falsos positivos, dado que presenta el mayor valor de población (aprox. 90 %) y el menor valor de energía de unión, difiriendo significativamente del resto de los complejos propuestos. Por otro lado, un análisis de su estructura muestra que la posición del ligando es muy cercana

al de la referencia (1.1 Å).

Una mirada detenida sobre las estructuras predichas por el método convencional (CADM) y el método modificada (WSBDM) se pueden observar en las figuras 5.4 C y D respectivamente en donde se muestra que para CADM la mejor predicción de la posición del anillo de la galactosa se encuentra correctamente ubicado en el sitio de unión, pero rotado 180 º; mientras que el complejo predicho por WSBDM coincide casi perfectamente con la estructura de referencia cristalográfica. En conclusión, los resultados muestran claramente que el WSBDM es capaz de predecir correctamente la estructura del complejo Gal-3:Gal, mientras que el CADM no.

5.3.4. Docking de N-Acetil-Glucosamina (GlcNAc) sobre Gal-3

De manera análoga al cálculo descripto en la sección previa, se utilizó el CADM y WSBDM para el ligando GlcNAc, observándose que los resultados obtenidos no son satisfactorios en ninguno de los dos casos. Por un lado, con el CADM no se observa un outlier de Población/ΔG_u que permita diferenciar el mejor complejo predicho por Autodock del resto de los complejos. El complejo con la más alta población (22 %) es cercano al de más baja energía y tiene un valor de RMSD respecto de la estructura de referencia de 5 Å. Mientras que el complejo con menor energía, que está más cercano a la estructura de referencia, no es fácilmente identificable. Por otro lado, el método modificado muestra resultados sutilmente mejores pero aún no satisfactorios. Por consiguiente, ambos métodos fallan para predecir correctamente el sitio de docking de GlcNAc dentro del sitio de reconocimiento de carbohidratos de Gal-3. La razón de está falla este posiblemente originada en el hecho de que GlcNAc tiende a ser ubicada cerca o más bien sobre el sitio que ocupa la galactosa, como es evidenciado a partir de que los valores de RMSD con respecto a la Gal, están aproximadamente en el rango 2.5-2.6 Å. Un análisis estructural sobre estos complejos propuestos, en conjunto con el detalle de los términos energéticos muestra que el término de vdW contribuye favorablemente a la energía de unión, originado principalmente por el contacto entre la cara hidrofóbica del carbohidrato con la cadena lateral del Trp68 (esto no resulta llamativo, ya que como se revela en base al trabajo de Guardia et al.,[18] este es un residuo clave para la interacción Galectina-carbohidrato). Adicionalmente, se observa en base al análisis de la estructura rayos-X de los complejos cristalizados con monosacáridos para otras galectinas (por ej.Gal-10), que la posición que toma el monosacárido en el complejo es tal que favorezca esta interacción hidrofóbica.

Esta observación nos condujo a repetir las simulaciones de docking agregando una restricción adicional de aproximadamente 2 kcal/mol al átomo de O4 de la GlcNAc en la posición del enlace glicosídico, con la galactosa ocupando el CBS en su posición cristalográfica. Los resultados que se desprenden de la figura 5.5 muestran que el CADM no es efectivo

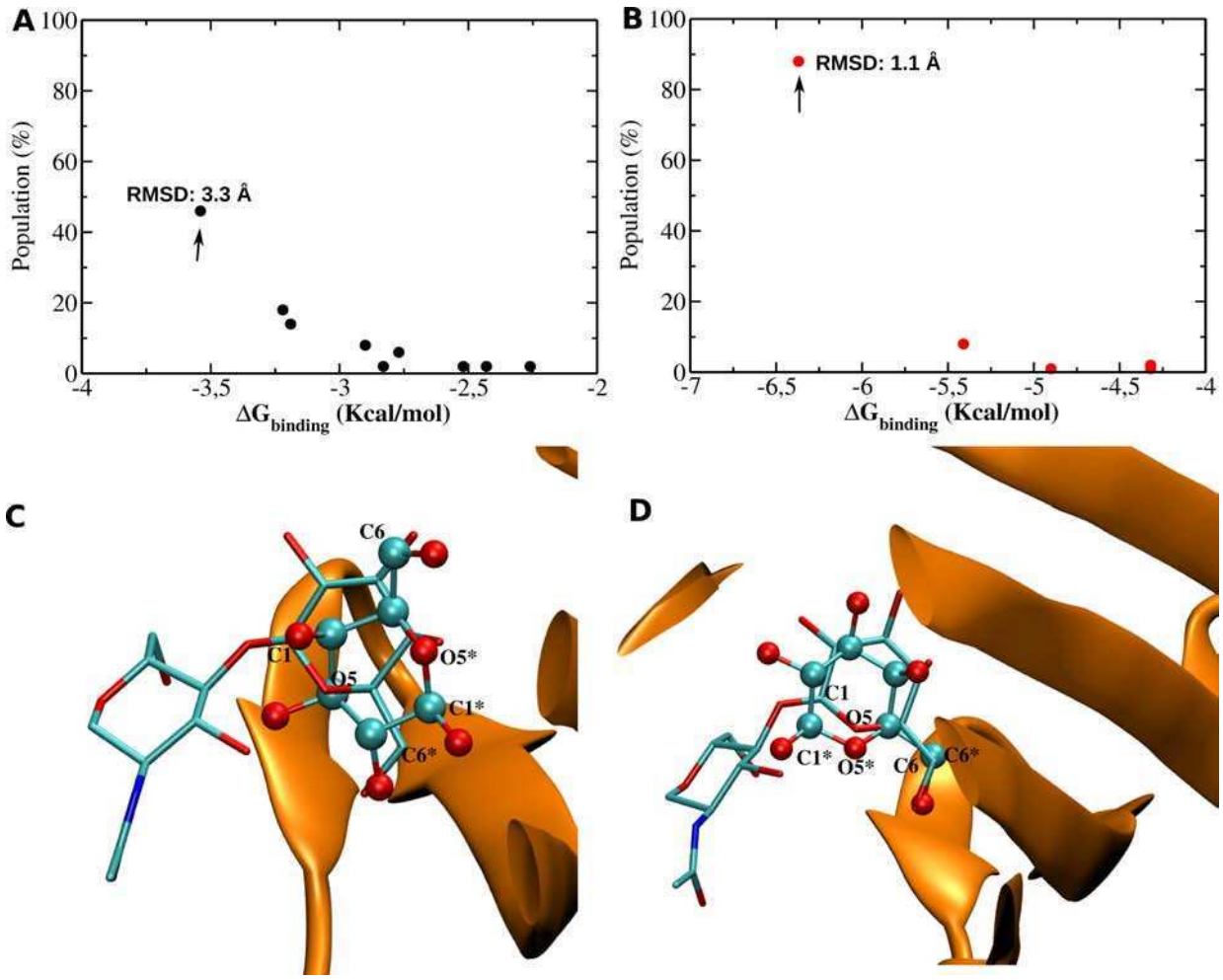

Figura 5.4. Resultado para el cálculo de docking de Gal sobre la estructura cristalográfica de Gal-3. Se muestra la población vs. ΔG_u utilizando **(A)** el método convencional de docking,CADM y **(B)** el método modificado, WSBDM. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomo de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1A3K). En los paneles inferiores se muestra: la estructura predicha para el complejo Gal3-Gal superpuesta sobre la estructura de referencia. **(C)** La estructura correspondiente al complejo de población más alta con el CADM predicha con un valor de RMSD de 3.3 Å. y **(D)** el segundo complejo predicho de más baja energía con un RMSD de 1.1 Å. La posición de los ligandos en los complejos predichos por el programa son mostrados en representación de "balls and sticks" mientras que la estructura de referencia se representa en "Licorice". Los átomos marcados con un asterisco (en las figuras C y D) corresponden a la estructura predicha, mientras que los que no están marcados con asterisco, a la estructura de referencia.

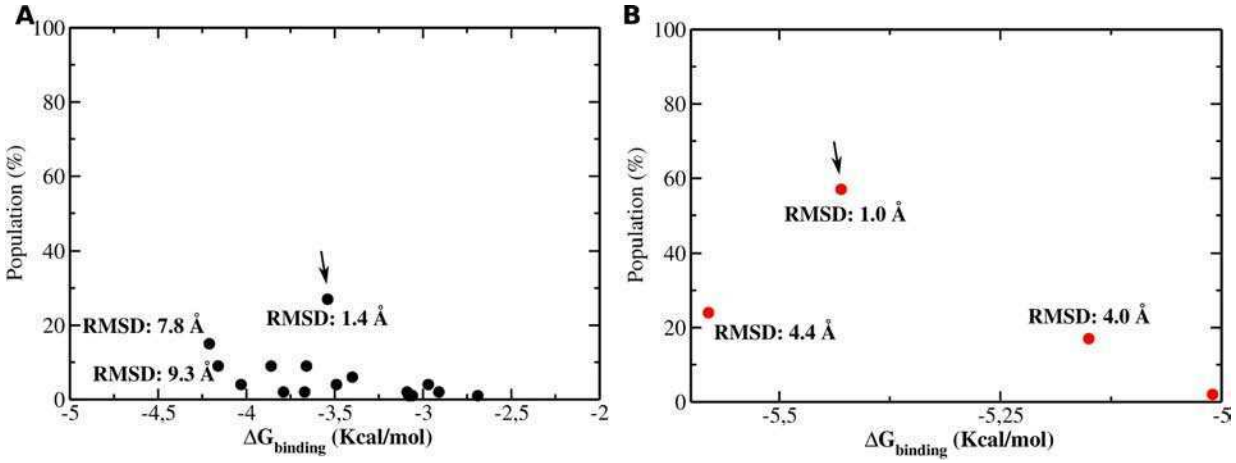

Figura 5.5. Población vs. ΔG_u para el cálculo de docking de GlcNAc sobre la estructura cristalográfica de Gal-3, con la galactosa unida en su sitio de unión y agregando una restricción armónica a la posición del O4 de GlcNAc para considerar el enlace glicosídico **(A)** utilizando el CADM y **(B)** y el método modificado (WSBDM). **(C)** Resultado para el cálculo de docking de N-acetil-lactosamina sobre una foto elegida al azar de la dinámica molecular del receptor de Gal-3 utilizando el CADM y **(D)** WSBDM. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1A3K).

al momento de posicionar correctamente al GlcNAc. Sin embargo, el método modificado muestra una mejor eficacia, identificando en este caso la estructura del complejo (RMSD de 1.0 Å) como el complejo de mayor población y ordenado en 2^0 lugar respecto de la energía, aunque con una diferencia menor a 0.5 kcal/mol en comparación con el complejo de mejor interacción.

5.3.5. Docking de N-Acetil-Lactosamina (LAcNAc) sobre Gal-3

Los resultados para el docking del disacárido LAcNAc sobre Gal-3, con el CADM y WSBDM utilizando los seis WS y la estructura cristalográfica del receptor en el complejo cristalográfico, así como conformaciones seleccionadas al azar a partir de un cálculo de dinámica molecular de la proteína libre para construir la grilla correspondiente al receptor, se muestran en la figura 5.6. Los resultados muestran claramente que el CADM (ver figura 5.6 A y C) es incapaz de dockear correctamente el disacárido, aún cuando un cálculo de redocking es realizado sobre una grilla construida sobre la estructura del receptor obtenida a partir de la estructura del complejo dilucidada por cristalografía de rayos-X. No se destaca ningún complejo y aquel de mayor valor de población o mejor energía de interacción (valores más negativos), muestran valores de RMSD >5 Å comparado contra el sistema de referencia. Las estructuras superpuestas (5.6 E) muestran que el CADM predice el complejo con la GlcNAc desplazada hacia el sitio de unión de Gal. Por otro lado, el WSBDM es

claramente capaz de predecir el complejo correcto (figura 5.6 B), y en donde la estructura correspondiente se superpone con la estructura de referencia. Como se muestra en la figura 5.6 F, la estructura predicha muestra un ajuste casi perfecto, ubicando ambos anillos de los azúcares correctamente y aún la cadena lateral N-acetilo con la orientación apropiada. Adicionalmente, también se probo el método pero eligiendo aleatoriamente conformaciones exploradas por la proteína libre para construir las grillas, observándose que también se obtiene una mejora en la capacidad predictiva del método, ya que como puede apreciarse en las figuras 5.6 C y D, a diferencia del CADM, para el WSBDM una estructura del complejo se destaca en el gráfico de población vs. energía, mostrando un valor de RMSD respecto de la estructura de referencia <1 Å.

La excelente eficiencia del WSBDM utilizando los 6 WS descriptos arriba, nos ha llevado a preguntarnos sobre la capacidad predictiva del método en relación al número de WS utilizados para modificar la grilla. De esta manera, basándonos en los resultados presentados en el capítulo 3, donde se muestra que los WS reemplazados por los -OHs del ligando tienen usualmente un alto valor de WFP y un bajo valor de R_{90}, surge una clara manera de seleccionar los WS para utilizar el WSBDM. La figura 5.7 muestra que los resultados para el cálculo de docking de LAcNAc sobre el CBS de Gal-3, incluyendo en el WSBDM un solo WS, aquellos de mayor valor de WFP, WS4 y WS2 (ver figura 5.6 A y B respectivamente), WS2 y WS4 juntos (ver figura 5.6 C) o tres correspondiendo a WS1, WS2 y WS4 (ver figura 5.6 D). Los resultados muestran que cuando se utiliza solamente un WS, los mismos son fuertemente dependientes de qué WS fue incluido para modificar la grilla convencional de Autodock4. Por ejemplo, cuando solo el WS2 es utilizado, el complejo de mayor población y menor energía no es una buena predicción del complejo correcto, apareciendo la misma en segundo lugar con el segundo valor de población más alto respecto de las otras estructuras. Por otro lado, cuando solo se utiliza el WS4, el complejo predicho con mayor población (ordenado en 2^0 lugar en términos de energía) es correctamente predicho (RMSD contra la referencia de solamente 0.5 Å). Utilizando ambos WS, los resultados mejoran significativamente, con un claro complejo que se destaca del resto (30 % en población y la mejor energía de interacción) con un valor de RMSD con respecto a la estructura de referencia de 0.6 Å.

Finalmente, si se utilizan tres WS, los resultados mejoran con respecto al uso de 2 WS, aunque no de manera significativa. Por consiguiente, estos resultados parecieran sugerir que al menos 1 WS por monosacárido es necesario incluir para producir una mejora considerable en el carácter predictivo del esquema convencional de docking. En resumen, para el presente caso, el WSBDM mejora de manera sustancial la calidad predictiva del método de docking y utilizar el disacárido completo parece ser una mejor estrategia que solo utilizar los monosacáridos por separado.

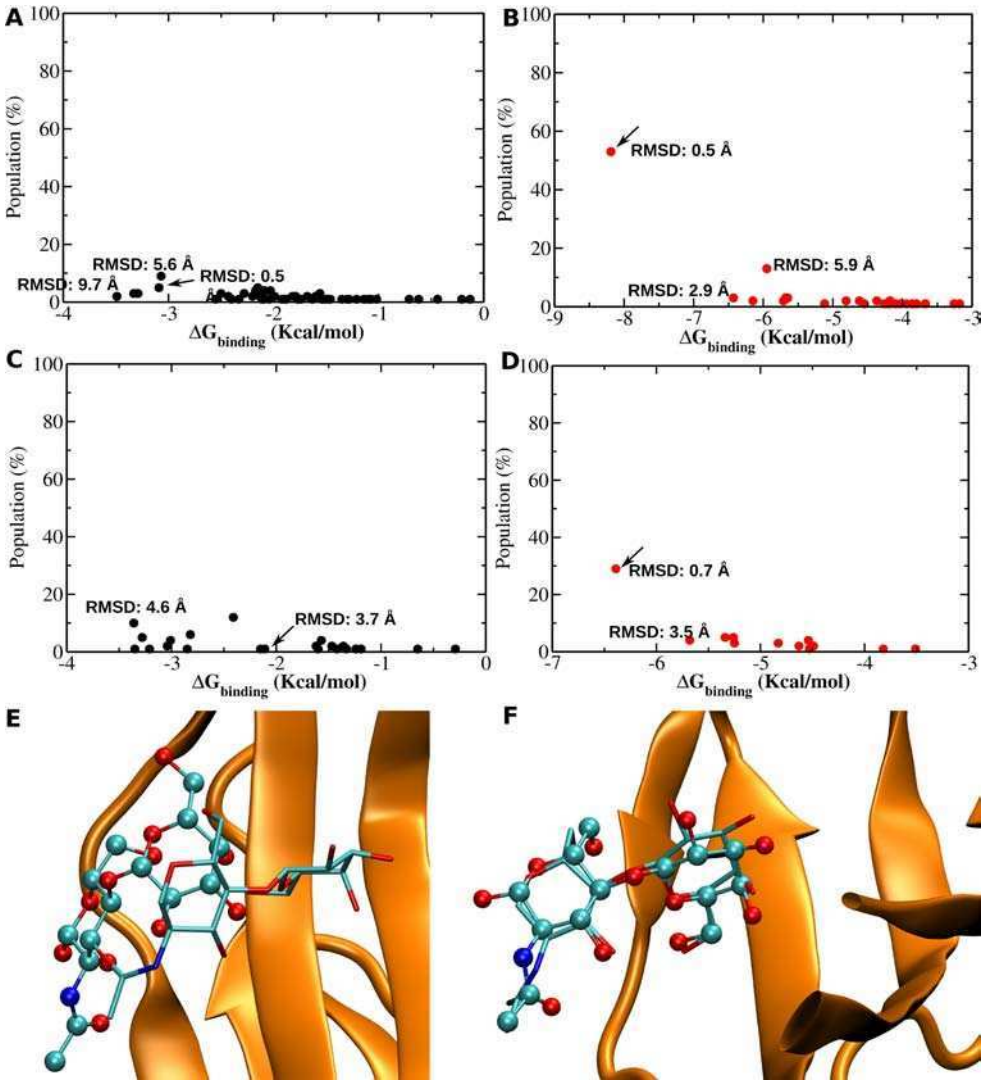

Figura 5.6. Resultado para el cálculo de docking de LAcNAc sobre la estructura cristalográfica de Gal-3 utilizando **(A)** el metódo convencional de docking (CADM) **(B)** y el modificado (WSBDM). **(C)** Resultado para el cálculo de docking de LAcNAc sobre una conformación elegida al azar de la dinámica molecular del receptor de Gal-3 utilizando el CADM y **(D)** WSBDM. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomo de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1A3K). En el panel inferiores, se muestra la estructura predicha para el complejo Gal3/N-acetil-lactosamina superpuesta sobre la estructura de referencia. **(E)** La estructura correspondiente al complejo de energía más baja con el CADM predicha con un valor de RMSD de 4.6 Å y **(F)** el complejo predicho de más baja energía para el WSBDM con un RMSD de 0.5 Å. La posición de los ligandos en los complejos predichos por el programa son mostrados en representación de “balls and sticks” mientras que la estructura de referencia se representa en “Licorice”.

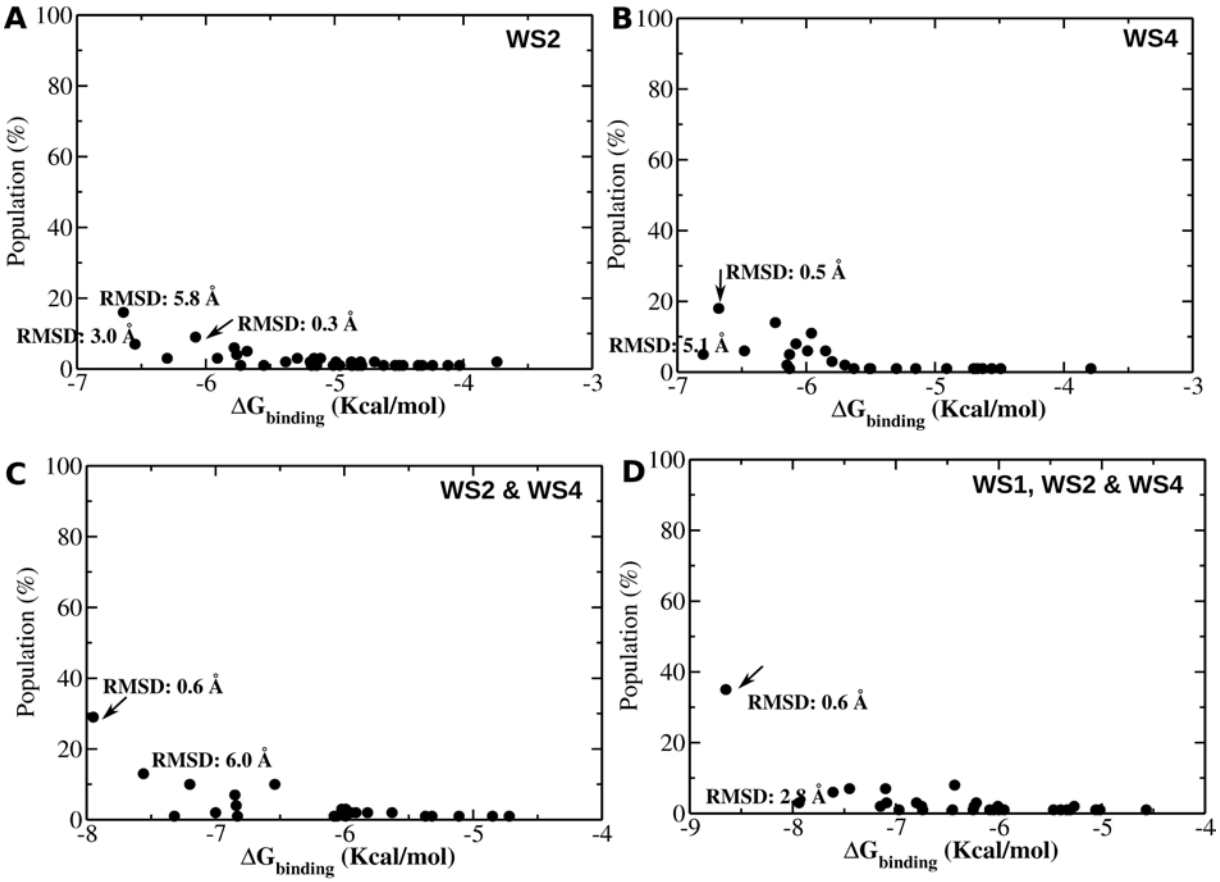

Figura 5.7. Resultado para el cálculo de docking de LAcNAc sobre la estructura cristalográfica de Gal-3 utilizando un número diferente de WS para construir la grilla modificada. En todos los casos se muestran los gráficos de Población vs. ΔG_u. **(A)** Solo utilizando el WS2; **(B)** solo el WS4, **(C)** Utilizando los WS2 y WS4 y en **(D)** Utilizando los WS1, WS2 y WS4. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1A3K).

5.3.6. Docking de trimanósido sobre Concanavalina A (Con-A)

Como se muestra previamente, Con-A presenta 11 WS con alto valor de WFP en el CBS. Tres de estos sitios de hidratación, WS7, WS8 y WS9, son reemplazados por los átomos O5, O6 y O3 de la primera manosa, los cuales establecen un total de 4 interacciones puente de hidrógeno con el receptor. El O2 de la segunda manosa claramente desplaza las aguas que ocupan el WS1 y el O4 las aguas que ocupan el WS5, mientras que el O4 desplaza el WS6 (ver tabla 5.1 y figura 3.4 A y B para más detalles). Sobre la base del estudio hecho en Gal-3 mostrado en la sección previa, se decidió utilizar directamente el trisacárido. En la figura 5.8 se muestran los resultados para el CADM y WSBDM del cálculo de docking del trimanósido pudiendo observarse que para el CADM, el cluster de mayor población y menor energía no da como producto el resultado correcto, ya que que el valor de RMSD respecto de la estructura de referencia es de 6.9 Å. Solamente el segundo cluster, de mayor población, el cuál también muestra un valor bajo de energía, predice una estructura más cercana al valor de referencia, con un valor de RMSD respecto de la referencia de aproximadamente 3.4 Å. Está estructura, mostrada en la figura 5.8 C, muestra al trimanósido considerablemente desplazado y cercano a la proteína, aunque esta correctamente orientado, con cada monosacárido se encuentra cercano a su sitio de unión. Nuevamente, como puede observarse en la figura 5.8 B, los resultados muestran una mejora significativa con el WSBDM en donde el complejo de energía más baja está claramente separado de los otros resultados y muestra un valor de RMSD de solamente 1.1 Å. Por otro lado, la figura 5.8 D muestra la estructura del complejo predicho, la cual es muy similar a la del complejo cristalográfico, con las dos manosas perfectamente ubicadas en el lugar correcto y la tercera levemente desplazada.

Del mismo modo que para el el caso de Gal-3, se utilizo Con-A para analizar la relación entre el número de WS utilizados para construir la grilla del WSBDM y la precisión de los resultados obtenidos. Para esto es importante elegir los WS de manera tal que nuestro conocimiento previo sobre la estructura del complejo no sesgue la elección de los mismos. La primera elección, como fue previamente mencionado, sería considerar aquellos WS con alto valor de WFP y bajo valor de R_{90}. Sin embargo, para CBS grandes, que unen tri o tetrasacáridos, la elección de los WS que cubren todo el CBS no parece ser una buena elección. Los datos mostrados previamente nos dicen que hay 5 WS con alto valor de WFP (10 veces mayor que en el seno del solvente o mayor) que pueden agruparse en dos conjuntos, cada uno de ellos en cada extremo del CBS. El primer grupo contempla a los WS7, WS8 y WS9, mientras que el otro conjunto incluye a los WS3 y WS6 respectivamente. Por consiguiente, se han realizado los cálculos de docking utilizando el WSBDM para el trimanósido construyendo grillas modificadas seleccionando un WS (él de más alta probabilidad de ambos conjuntos, por ejemplo WS8 o WS3) ó 2 WS (combinando ambos) o

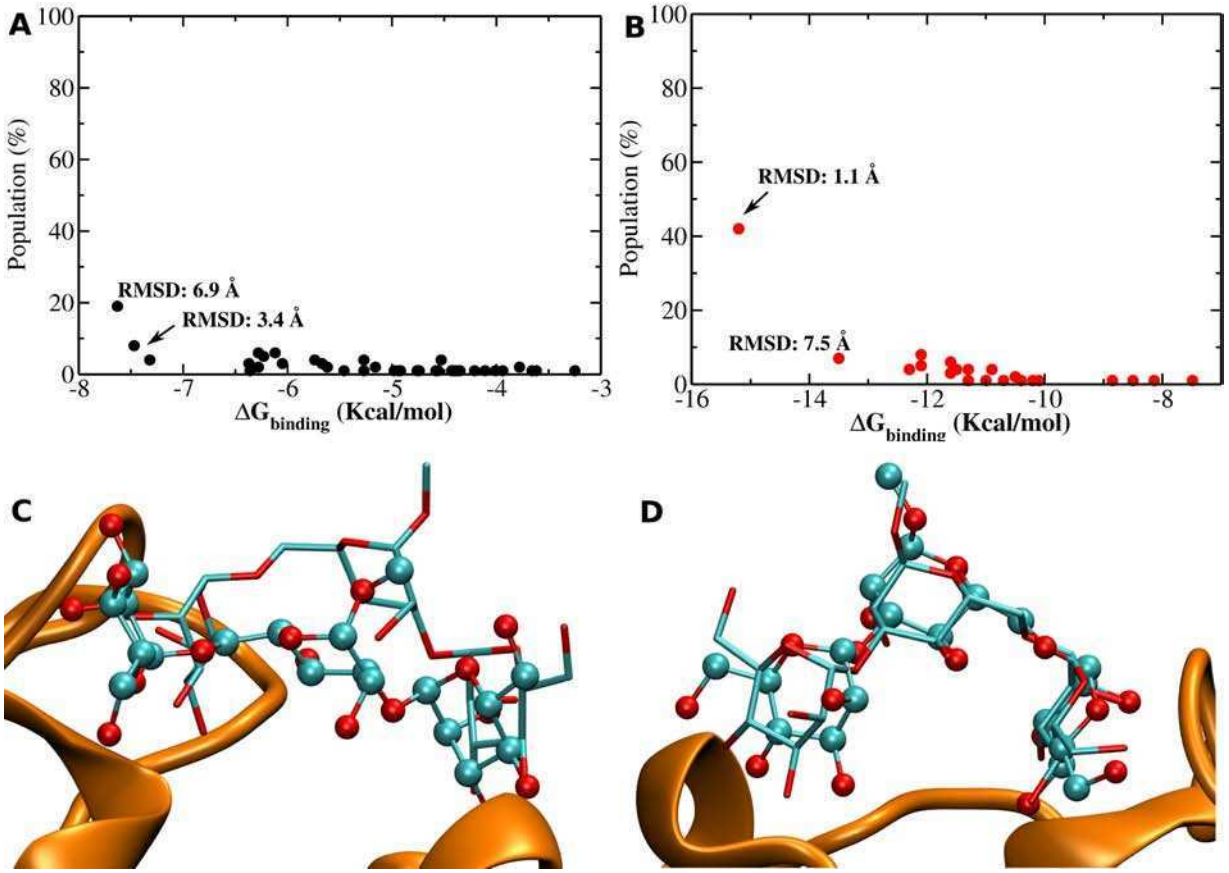

Figura 5.8. Resultado para el cálculo de docking del trimanosido sobre la estructura cristalográfica de Con-A. Se muestra la población vs. ΔG_u utilizando **(A)** el método convencional de docking (CADM) y **(B)** el método modificado (WSBDM). Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1ONA). En los paneles inferiores, se muestra la estructura predicha para el complejo del ConA-trimanósido superpuesta sobre la estructura de referencia. **(C)** La estructura correspondiente al mejor complejo (en términos de RMSD) con el CADM predicha con un valor de RMSD de 3.4 Å. y **(D)** el complejo mejor rankeado para el WSBDM con un RMSD de 1.1 Å. La posición de los ligandos en los complejos predichos por el programa son mostrados en representación de "balls and sticks" mientras que la estructura de referencia se representa en "Licorice".

3 WS (seleccionando el mejor WS de cada grupo, WS3, WS8 y WS1) y también los 5 WS de más alto valor de WFP. Los resultados muestran que cuando se utiliza únicamente un WS, el método aún es capaz de predecir correctamente la estructura del complejo, pero este resultado depende fuertemente del WS seleccionado. Si solo se utiliza el WS8, el complejo de menor energía muestra un valor de RMSD contra la referencia muy bajo (RMSD 0.9 Å), pero si se utiliza únicamente el WS3, los resultados finales no son satisfactorios (ver figura 5.9 A y B). Los resultados con 2 y 3 WS son claramente mejores que los obtenidos con el CADM, pero cuando se utilizan solo 2 WS el mejor complejo tiene un valor alto de RMSD respecto de la estructura de referencia, comparado con el caso donde solamente el WS8 es utilizado para construir la grilla (ver figura 5.9 C). Con tres WS, aunque el complejo de menor energía es una mala predicción del complejo correcto, el segundo complejo de menor energía corresponde al complejo correcto (ver figura 5.9 D). Finalmente, cuando se utilizan los 5 mejores WS (figura 5.9 D), los resultados son similares (aún en términos de RMSD) a aquellos obtenidos con todos los WS, con el complejo de menor energía claramente separado del resto en términos de energía/población y con un bajo valor de RMSD con respecto a la referencia cristalográfica (RMSD de 2.1 Å).

En conjunto, estos resultados sugieren que cuando solo unos pocos WS son utilizados, el carácter predictivo del método varia significativamente. Esto no es sorprendente, dado que no todos los WS son reemplazados par los grupos -OHs del ligando. Por consiguiente, parece ser que la mejor elección es utilizar muchos WS, al menos uno por monosacárido. Por último, del mismo modo que para Gal-3, se ha repetido el cálculo de docking en las mismas condiciones pero utilizando distintas conformaciones del receptor obtenidos a partir de la simulación de dinámica molecular de la proteína libre. Los resultados muestran que el CADM es incapaz de predecir correctamente la estructura del complejo. En cambio, los resultados obtenidos para el WSBDM son mejores, pero no tan buenos como los obtenidos para el receptor en la conformación aportada por la estructura cristalográfica del complejo.

5.3.7. Docking de maltotriosa sobre la proteína surfactante D (SP-D)

Como otro caso de prueba para trisacáridos, se ha seleccionado a la proteína SP-D. Esta proteína es un lectina de tipo C que presenta un dominio de unión a carbohidratos, capaz de unir maltotriosa (ver Figura 5.2 F). En ella se han identificado 5 WS en el CBS (tabla 5.1), dos de los cuales muestran un alto valor de WFP. Los resultados para el docking de maltotriosa al CBS de SP-D utilizando tanto el CADM como el WSBDM muestran que para el CADM, el complejo de menor energía es predicho lejos del sitio de unión (RMSD 4.1 Å respecto de la estructura de referencia), con solo uno de los monosacáridos (número 3) correctamente ubicado. Muy cercano en energía y población a está estructura, se encuentra el segundo complejo el cuál está mucho más cerca de la estructura de referencia (RMSD

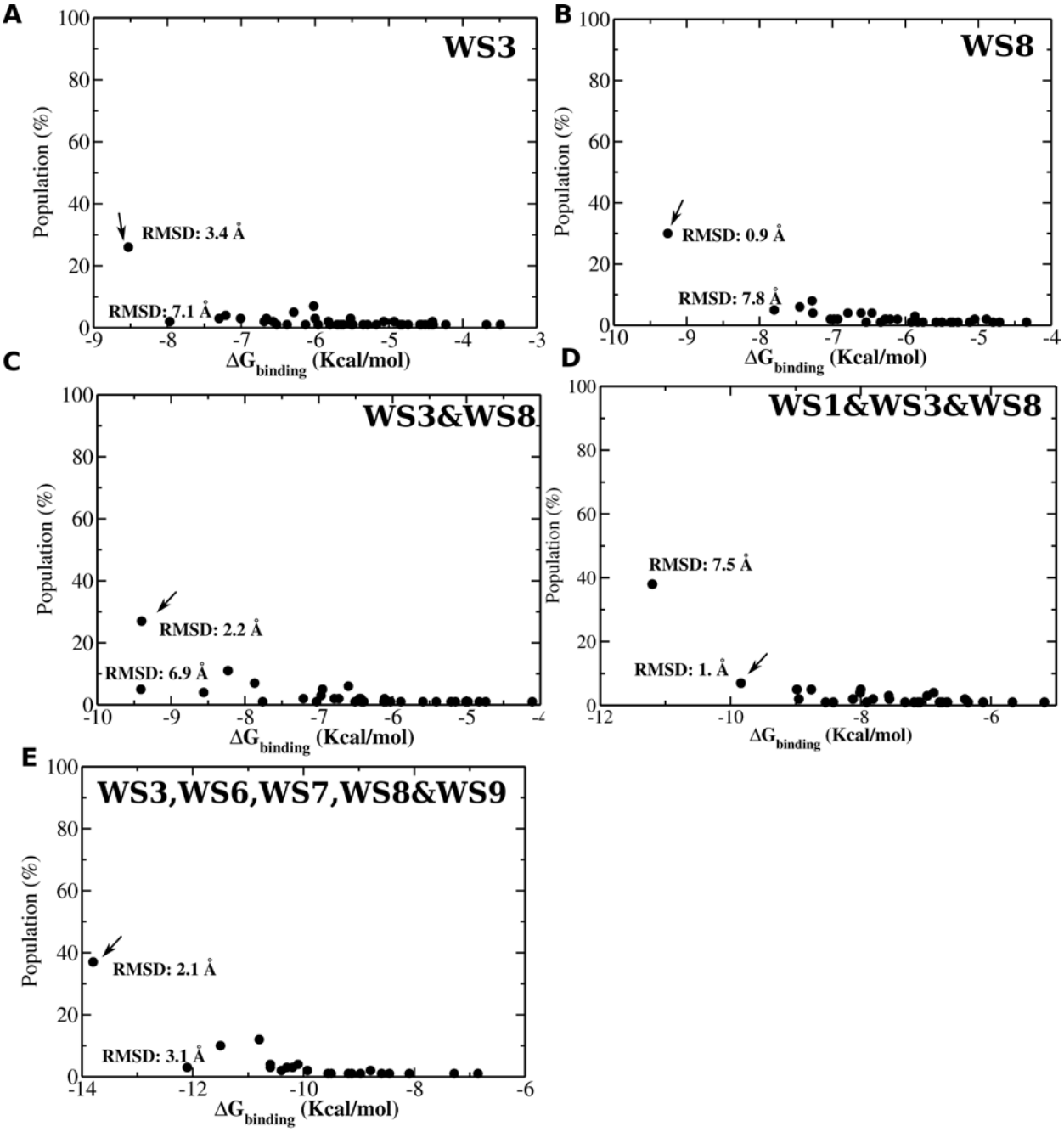

Figura 5.9. Resultado para el cálculo de docking del trimanósido sobre la estructura cristalográfica de Con-A utilizando un número diferente de WS para construir la grilla modificada. En todos los casos se muestran los gráficos de Población vs. ΔG_u. **(A)** Solo utilizando el WS3; **(B)** el WS8, **(C)** Utilizando el WS3 y el WS8 y **(D)** utilizando el WS1, WS3 y WS8. Finalmente, se muestra en el gráfico **(E)** los resultados a partir de la grilla construida utilizando los WS3, WS6, WS7, WS8 y WS9. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 1A3K).

de 1.8 Å) y el cual ubica los monómeros correctamente. Por otro lado, el WSBDM, ordena claramente el complejo que esta mas cercano a la estructura de referencia (con un RMSD de 1.8 Å) en primer lugar, con un valor de energía de unión significativamente menor y un valor de población mayor que el resto de los complejos predichos. Una inspección visual de los resultados obtenidos en relación a la estructura de referencia, muestra claramente que los tres monosacáridos que conforman el ligando se encuentran correctamente posicionados y orientados en sus respectivos sitios de unión, con el tercero mostrando una coincidencia perfecta mientras que los otros dos levemente corridos. Claramente, el WSBDM permite una clara predicción de la estructura del complejo.

5.3.8. Docking de Hialuronato (HA_4) sobre CD44

Como caso final de estudio se ha seleccionado CD44, un receptor presente la superficie celular que une ácido hialurónico. La estructura cristalina del complejo muestra un dominio de reconocimiento de carbohidratos que une un octasacárido. Sin embargo, solamente 4 monosacáridos están en contacto directo con la proteína y por consiguiente, fueron utilizados como ligandos para el cálculo de docking, resultando en dos unidades disacáridicas repetidas: ácido β-D-Glucoronico y β-D-N-Acetil-Glucosamina, unidos a través de enlaces glicosídicos alternados β-1,4 y β-1,3 (ver figura 5.1 D). Además, dado que tanto la estructura del complejo y del receptor libre (apo-proteína) se encuentran resueltas por difracción de rayos-X y depositadas en el *p*rotein data bank, se ha probado la variabilidad de los resultados para estos dos casos.

Para realizar el WSBDM, como en el caso previo, primero se determino el mapa de solvatación del CBS. La simulación de dinámica molecular de la apo-proteína en solvente explicito muestra que el CBS contiene 5 WS (como se puede observar en la tabla 5.1) de los cuales, dos presentan un alto valor de WFP y otros dos con un valor intermedio (aproximadamente 5 veces mayor que en el seno de la solución).

Los resultados para el CADM muestran que hay una moderada variabilidad de los mismos dependiendo de la estructura elegida para realizar el cálculo de docking. El docking de HA_4 tomando al receptor sobre la estructura cristalográfica del complejo, permite una correcta identificación de la posición correcta del ligando (RMSD de solamente 1.0 Å) con la energía más baja y el valor de población más alta (alrededor del 40 %). No obstante, si se utiliza la estructura del receptor cristalizada sin ligando, es interesante notar que el complejo predicho de más baja energía muestra al ligando posicionado al revés y con un valor de RMSD de 11.4 Å, prediciendo a la estructura correcta en segundo lugar (ver figura 5.10 A y C).El WSBDM, por su parte, muestra una mejora significativa para ambos casos, dando como resultado los complejos de más alta población para las estructuras mejor rankeadas y con menor valor de RMSD contra la referencia cristalográfica. Aún para el

cálculo de docking de la apo-estructura, el WSBDM ordena la estructura correcta en primer lugar (ver figura 5.10 B) y la estructura predicha con el ligando al revés está posicionada en segundo lugar mostrando también un valor de población alto. La estructura mejor ordenada para el WSBDM se muestra en la figura 5.10 D donde es posible observar una superposición casi perfecta con la estructura cristalográfica, para el tetrasacárido completo.

Finalmente, se evaluó esta estrategia sobre conformaciones tomadas de manera aleatoria desde la dinámica molecular de la apo-proteína, observándose una correcta predicción de la estructura del complejo con el WSBDM, mientras que el CADM sigue siendo incapaz de predecir la estructura correcta.

5.4. Discusión

El propósito del presente capítulo fue analizar si la información de la estructura del solvente en el CBS, derivada a partir de simulaciones de dinámica molecular en solvente explícito y descripta mediante la caracterización de los WS, podría ser utilizada para mejorar la eficiencia de los esquemas de docking convencionales, utilizados para la predicción de la estructura de los complejos proteína-carbohidrato. Con este propósito, comparamos la calidad de los resultados obtenidos utilizando el método de docking convencional (CADM) con aquel en el que se ha incluido la información del mapa de solvatación (WSBDM), en cuanto a su capacidad de predecir correctamente la estructura de los complejos, y además de poder identificar la misma de los numerosos falsos positivos que se arrojan como posibles resultados, utilizando para ello diferentes proteínas que unen carbohidratos y con ligandos que varían en tamaño desde monosacáridos hasta tetrasacáridos.

En total, se realizaron más de 30 cálculos de docking en donde se vario el tamaño del los ligandos, la estructura del receptor y el número de WS utilizados para modificar el esquema de docking convencional. A continuación se presenta un análisis global de los resultados:

5.4.1. Análisis general de los resultados obtenidos

Un resumen de todos los resultados obtenidos en el presente capitulo se condensan en la figura 5.11. Por un lado, la figura 5.11 A muestra los valores de RMSD del ligando respecto de la referencia cristalográfica para el complejo de menor energía predicho por el CADM (columnas negras) y el WSBDM (columnas rojas). En segundo lugar, se gráfico el valor de RMSD para el complejo predicho con menor valor de RMSD entre todos los clusters, con CADM (columnas grises) y WSBDM (columnas naranjas). De este modo, los resultados mostrados en la figura 5.11 A, muestran que existe una clara diferencia en la capacidad predictiva entre los dos métodos. Mientras que aquellos complejos que el CADM ordena primero (columnas negras) representan una estructura incorrecta (RMSD mayor a

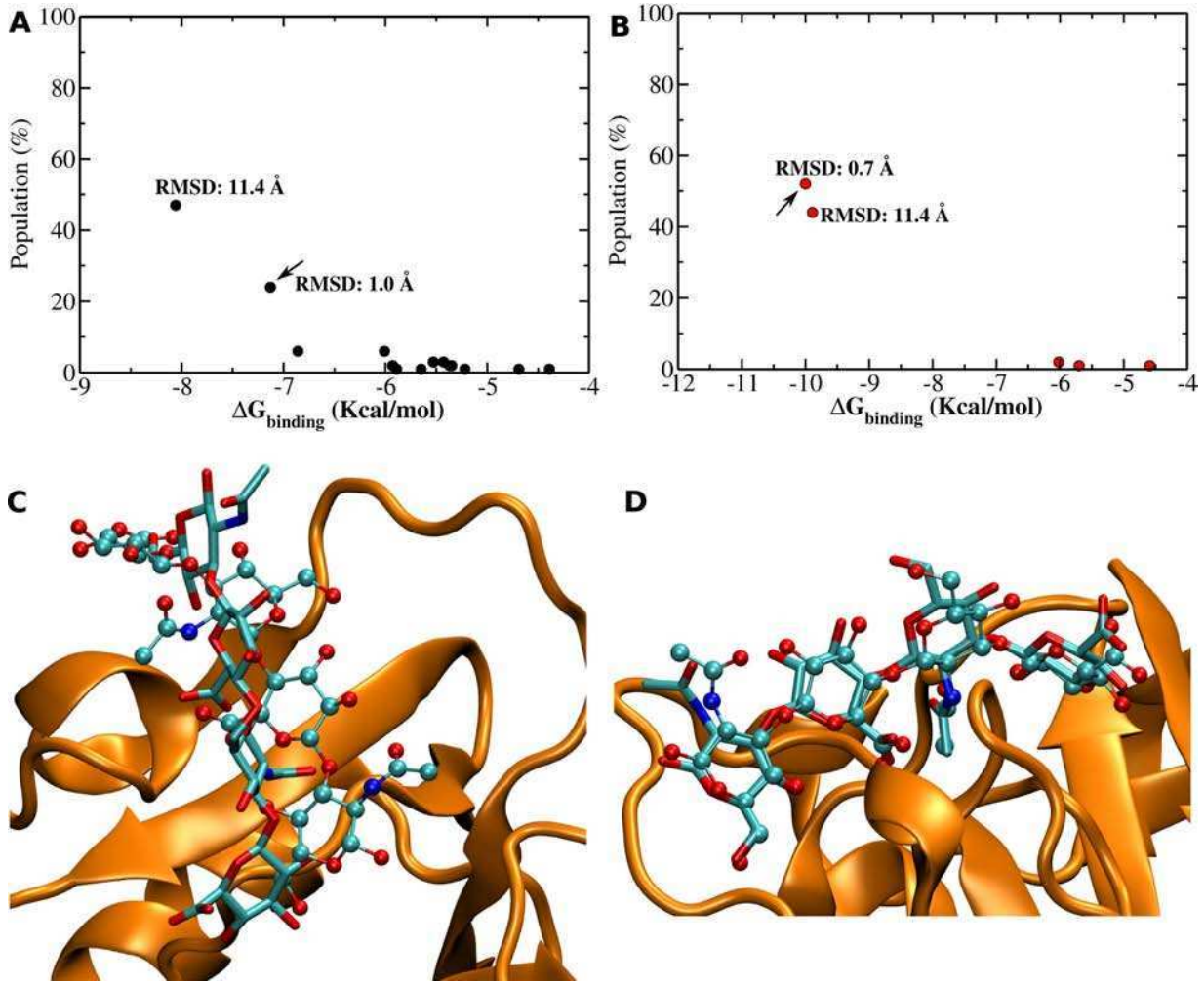

Figura 5.10. Resultado para el cálculo de docking del tetrasacárido hialuronato sobre la estructura cristalográfica de CD44. Se muestra la población vs. ΔG_u utilizando **(A)** el método convencional de docking (CADM) y **(B)** el método modificado (WSBDM). Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD teniendo en cuenta los átomos de carbono entre el complejo predicho por el programa de docking y la estructura cristalográfica del complejo de referencia (PDB ID 2JCP). En los paneles inferiores, se muestra la estructura predicha para el complejo del CD44-hialuronato superpuesta sobre la estructura de referencia. **(C)** La estructura correspondiente al complejo mejor rankeado con el CADM predicha con un valor de RMSD de 11.4 Å. y **(D)** el complejo mejor rankeado con un RMSD de 0.7 Å. La posición de los ligandos en los complejos predichos por el programa son mostrados en representación de "balls and sticks" mientras que la estructura de referencia se representa en "Licorice".

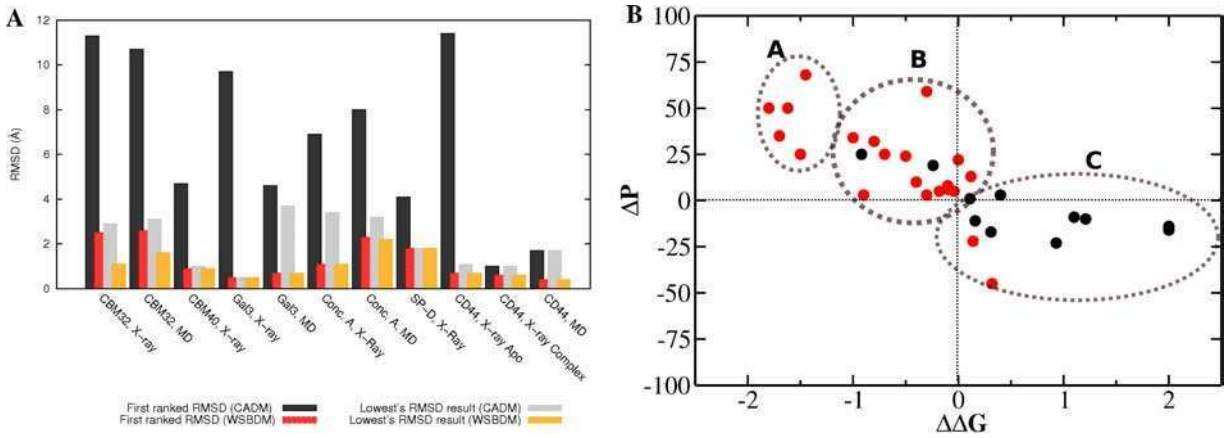

Figura 5.11. Para todos los cálculos de docking realizado se muestra **(A)** El valor de RMSD para el complejo predicho de más baja energía, es decir el mejor rankeado y para el de menor valor de RMSD para el CADM en columnas negras y grises respectivamente. Para el WSBDM, se muestran lo mismo, es decir el valor de RMSD complejo predicho de más baja energía y para el de menor valor de RMSD pero en columnas rojas y naranjas respectivamente. En la figura **(B)**, se muestra el gráfico de $\Delta\Delta G_u$ vs. ΔPop para los cálculos de docking realizados con el CADM (puntos negros) y el WSBDM (puntos rojos). Ver el texto para la definición de $\Delta\Delta G_B$ y ΔPop

2 Å), los complejos ordenados en primer lugar con el WSBDM (columna roja) están cerca (RMSD entre 2 Å y 3 Å) o muy cerca (menor a 1 Å) respecto de la estructura de referencia. Además, para la mayor parte de los casos, en el WSBDM los complejos que tienen el menor valor de RMSD con respecto a la estructura de referencia, usualmente tienen el menor valor de energía (comparar columnas rojas con columnas naranjas en la figura 5.11 A). Es interesante notar que si se observa el complejo mejor posicionado en el ranking, no es en la mayoría de los casos el de menor valor de RMSD para el CADM , mostrando los mismos amplias diferencias (comparar columnas negras y grises para el CADM). No obstante, para el WSBDM, existe una buena correlación (comparar columnas rojas y naranjas). Por consiguiente, la principal dificultad de Autodock4 no parece estar en su capacidad de encontrar el complejo correcto, pero si en su capacidad de rankear correctamente al mismo entre las numerosas posibilidades generadas por el algoritmo genético, de acuerdo al valor de energía libre de unión, un hecho que también es observado en otros trabajos.[4,6,19,20] Esta observación es consistente con los resultados generales presentados en este trabajo, los cuales muestran que la modificación a la función de puntuación de Autodock4, que determina la energía libre de unión, genera una mejora significativa en los resultados obtenidos.

Por otro lado, la figura 5.11 B pone énfasis en poder analizar la capacidad predictiva del método en relación a su precisión. Para ello, se ha calculado la diferencia en la energía de unión predicha ($\Delta\Delta G$) y la diferencia en la población (ΔP) para el mejor complejo (aquel con el menor RMSD) y el mejor rankeado de los complejos predichos, es decir, aquel que

presenta menor valor de energía de unión. En caso de que este sea además el que presenta menor valor de RMSD con respecto a la referencia cristalográfica, se toma la diferencia con el segundo complejo de menor energía. Por consiguiente, un valor de $\Delta\Delta G$ menor a cero, implica que el complejo de menor valor de RMSD tiene un valor de energía libre de unión más bajo que el resto de los complejos predichos y la magnitud del valor de $\Delta\Delta G$ refleja la diferencia entre el complejo mejor predicho y el primer falso positivo. Por otro lado, un valor positivo de $\Delta\Delta G$ significa que el mejor complejo tiene una energía de interacción más alta respecto de aquel rankeado primero indicando que la mejor predicción es incorrecta. De manera análoga, un valor de ΔP positivo, significa que la mejor predicción tiene un valor de población mayor comparado con los otros complejos predichos, mientras que un valor negativo de ΔP significa que la mejor predicción tiene un valor de población más bajo con respecto a las otras predicciones incorrectas. De este modo, los resultados localizados en la esquina superior izquierda del gráfico mostrado en la figura 5.11 B corresponden a aquellos casos en los que el complejo correcto es rankeado correctamente (es decir, tiene el menor valor de energía de unión y el mayor valor de población) y por lo tanto ubica correctamente al ligando dentro del CBS (RMSD menor a 1 Å respecto de la referencia) correspondiendo estos cálculos a resultados exitosos, en el sentido que los mismos podrán predecir correctamente la estructura del complejo proteína-carbohidrato.

Una inspección general del gráfico presentado en la figura 5.11 B muestra indudablemente que el WSBDM (puntos rojos) tiene una mejor eficiencia en cuanto a predecir el complejo correcto según los datos de cristalografía que el CADM (puntos negros), con la mayor parte de los resultados localizados en la esquina superior izquierda (grupos A y B). Únicamente dos resultados computados con el WSBDM se ubican en el cuadrante inferior-derecho del gráfico (grupo C), correspondiendo al caso de CBM32 utilizando tanto la estructura del receptor aportada para la estructura cristalográfica del complejo o alguna elegida a partir de la simulación de dinámica molecular para construir la grilla. Un análisis más detallado de los resultados obtenidos para el WSBDM nos permite identificar un primer grupo de resultados (grupo A) donde el complejo correcto es predicho con al menos 1 kcal/mol por debajo del resto de los complejos y con un valor de ΔP mayor al 25 %, lo cual muestra claramente que los complejos correctos son fácilmente identificables del resto de falsos positivos dado que los mismos se comportan como outliers en la distribución de datos que representan el total de los complejos predichos.

Todos los resultados que pueden ser agrupados dentro de este conjunto corresponden además, a aquellos escenarios donde se hizo redocking, es decir que la estructura del receptor que se utilizo para construir las grillas es aquella que estaba previamente cristalizada con el ligando. El grupo B incluye a la mayoría de los resultados obtenidos por el WSBDM cuando la estructura del complejos fue tomada a partir de una simulación por dinámica

molecular, un resultado que no es del todo inesperado. Para este caso, la diferencia en la energía libre de unión y el valor de población es pequeño. Finalmente, es de destacar que el único resultado generado por el CADM que pertenece a este grupo, corresponde al cálculo de docking de HA_4 sobre CD44, tomando la estructura cristalográfica del complejos para construir la grilla.

5.4.2. Efecto del tamaño del ligando

Cuando se comparan los resultados en su conjunto en relación al tamaño de ligando, **la correcta predicción de la unión de los monosacáridos parece ser más difícil que para carbohidratos más largos**. Esto se evidencia en CBM32 cuando aún el WSBDM falla en ordenando correctamente la estructura correspondiente a lo revelado por los datos experimentales aportados por cristalografía de rayos-X. Además, en el caso de Gal-3, se observa que utilizar el disacárido es claramente una mejor opción que utilizar los monosacáridos por separado. Por otro lado, los resultados de docking para los casos en los que se utiliza tri y tetra-sacáridos son los más precisos, ya que los complejos predichos tienen un bajo valor de RMSD con respecto a la referencia y además el complejo correcto rankea primero con respecto a la energía. Por consiguiente, también para este caso, utilizar azúcares largos parece ser una mejor idea que utilizar los monosacáridos individualmente.

Sin embargo, es importante notar, que este no siempre es el caso, en particular cuando se incrementa el tamaño del azúcar, **dado que el algoritmo genético que Autodock utiliza para explorar todas las posibles configuraciones del ligando en el sitio de unión comienza a fallar cuando el numero de grados de libertad torsionales del ligando se incrementan** (en particular, con un número de grados de libertad torsionales mayor a 8-10, la exploración del espacio configuracional ya no es tan eficiente). Teniendo en cuenta que **cada monosacárido agrega al menos 2 grados de libertad torsionales**, para tetrasacáridos o ligandos más grandes, dockear pequeñas partes del hidratos de carbono (di- o tri-sacáridos), con una restricción armónica de algunas kcal/mol sobre los átomos de oxígeno del azúcar que están involucrados en la unión glicosídica, parece ser una mejor estrategia que utilizar el carbohidrato completo.

5.4.3. Efecto del número y característica de los WS utilizados

Si consideramos ahora el número de WS a utilizar, los resultados muestran que el uso de todos los WS parece ser una mejor decisión que utilizar solo algunos. Y aunque es posible predecir correctamente la estructura del complejo con solamente un WS, los resultados son muy dependientes de qué WS se seleccione. Sin embargo, con el objetivo de determinar cuántos WS deben ser determinados, caracterizados y utilizados en el WSBDM para un

caso particular, como **regla general podría decirse que se deben utilizar 1 o 2 WS por subunidad monosacarídica. Sin embargo, es importante elegirlos de manera tal que estén extendidos sobre todos el CBS**.

Por último, se debe destacar que no resulta llamativo que se observen mejoras significativas si solo se eligen aquellos WS con un alto valor de WFP, con una sutil mejora si se eligen aquellos de bajo valor de WFP, dado que la función que introduce el modificación de las funciones de puntuación originales de Autodock4 son pesadas por el valor de WFP (ver ecuación 2.11)

5.4.4. Cómo utilizar el WSBDM

En está sección describiré brevemente como aplicar el WSBDM para un problema en particular comenzando desde la estructura del receptor y el ligando por separado.

El protocolo puede ser planteado en 4 pasos: **(1)** Calcular WS (lo que puede hacerse de manera automática con el programa de clusterización de aguas que se describe en el capítulo 2; **(2)** seleccionar los WS para modificar la grilla que se construye a partir del receptor; **(3)** realizar el WSBDM y **(4)** analizar los datos.

1. En primer lugar, la estructura del receptor debe ser sometida a una simulación de dinámica molecular en aguas explícitas alrededor de 20-40 ns. Desde la simulación, los WS adyacentes al sitio de reconocimiento de carbohidratos (CBS) de la proteína deben ser determinados y caracterizados.

2. Los sitios de hidratación (WS) deben ser analizados y aquellos que tienen un valor alto de WFP (usualmente mayor a 5) deben ser seleccionados. Los WS elegidos deben estar bien distribuidos sobre toda la extensión del CBS y el número de WS debiera estar en el rango de 1-2 WS por monosacárido. Las grillas debieran ser construidas sobre diferentes estructuras del receptor de la dinámica molecular de la proteína libre, y finalmente todos esos resultados debieran ser analizados en su conjunto seleccionando aquellos complejos de menor energía y mayor población

3. Para cada grilla, entre 100-200 cálculos de docking individuales deben ser realizados y clusterizados tal como se describe en el capitulo 2 .

4. Para analizar los datos, se deben construir los gráficos de población vs. ΔG_u, prestando particular atención a aquellos puntos que son outliers, los cuáles tienen un valor significativamente menor de energía de unión y mayor de población que el resto de los otros resultados ($\Delta\Delta G_B$ >1 kcal/mol y ΔPop >25 %). Si no se observan cluster outliers, un número diferente de fotos o WS debieran ser utilizados para reconstruir las mapas de energía. Un complejo confiable debiera aparecer como el mejor rankeado

utilizando diferentes configuraciones del receptor, y una vez que ha sido encontrado, este debiera mantenerse como el mejor rankeado cuando se incrementa el número de WS que se utilizan para reconstruir el mapa de energía.

Como una observación final, es importante discutir el costo computacional requerido para utilizar el WSBDM respecto del CADM. Una vez que los mapas de energía fueron construidos, realizar el cálculo de docking lleva la misma cantidad de tiempo con ambos métodos. Sin embargo, mientras que construir la grillas con el método de docking convencional requiere solamente una única estructura del receptor, para construir el mapa de energía modificado por los WS, debe realizarse y analizarse una simulación de dinámica molecular. No obstante, actualmente, el tiempo de computo requerido para realizar una simulación de dinámica molecular no es un limitante, requiriéndose aproximadamente para una proteína de tamaño medio como Gal-3, CBM40 o Con-A aproximadamente 8 hs por cada ns en un CPU de 8 núcleos, y por consiguiente utilizando 32 núcleos, en donde se ha visto que el código de AMBER escala linealmente, se toma menos de 1 semana en realizar una simulación de DM de alrededor de 50 ns. Finalmente, es interesante comentar que con la implementación de los algoritmos de simulación de dinámica molecular en GPU (graphic processing unit), este tiempo de computo se reduce un par de horas. Finalmente, la determinación del mapa de solvatación del CBS, el cálculo de WFP y R_{90} lleva entre 10 minutos y 1 hora (dependiendo el tamaño del sistema a analizar), con el programa de clusterización de aguas desarrollado en la presente tesis y descripto previamente en el capítulo 2

5.5. Metodología específica del capítulo

Para realizar los cálculos de docking se utilizo el programa Autodock4.[21] De manera resumida, el protocolo empleado para los cálculos de docking es como se detalla a continuación: En primer lugar, el programa construye una mapa de energía para cada tipo de átomo del ligando, basado en una única conformación del receptor, en donde las interacciones de no unión proteína-ligando (que incluyen solamente las contribuciones electrostáticas y de vdW) son calculadas. Por consiguiente, durante el cálculo de docking, la energía de interacción proteína-ligando es estimada para cada posición/conformación del ligando de manera directa sobre la grilla.

En segundo lugar, se genera un conjunto inicial de posiciones/conformaciones del ligando las cuales son ubicadas de manera aleatoria sobre la grilla y para cada una de ellas se computa la energía de interacción. De este modo, las conformaciones malas mostrarán un valor de energía de interacción pobre y serán eliminadas, mientras que las mejores se conservan. Luego, se generan nuevas estructuras de docking en base a está selección

en función de los grados de libertad estructurales del ligando seguido de un paso de minimización de la energía. Este tipo de exploración en base a un algoritmo genético de tipo Lamarckiano con búsqueda local (esto último implica un sutil cambio conformacional del ligando en pos de minimizar de la energía de interacción proteína-ligando) se continua hasta que la mejor conformación (aquella de menor energía) es alcanzada, correspondiendo al complejo proteína-ligando propuesto. Este procedimiento se conoce como corrida de docking. Usualmente para cada par proteína-ligando, se realizan del orden de cientos de corridas de docking y los resultados son clusterizados de acuerdo a la posición/conformación del ligando, dando como resultado un parámetro conocido como población, el cual se informa para cada complejo propuesto por el programa (en otras palabras, la población viene a ser la cantidad de veces en el que una corrida de docking individual alcanza un determinado modo de unión en la grilla utilizada).[1,2,17,21] Es interesante resaltar en este punto, que los resultados que informa Autodock, están ordenados desde el de menor energía libre de unión de manera decreciente (donde se muestra de manera directa la estructura del complejo que une con mayor afinidad dentro de cada conjunto representativo) y muchas veces, de manera contraria a lo esperado, el complejo de menor energía no resulta el de mayor población. Sin embargo, en la mayor parte de los casos, la diferencia de energía entre el mejor rankeado y el de mayor población (en caso que no sea también el mejor rankeado) es pequeña. Está diferencia, está debida a que Autodock utiliza un método pseudo-estocástico para explorar el espacio de conformacional y por otro una estrategia determinista en base al computo de la energía libre en función del campo de fuerza utilizado.

Para los cálculos de docking planteados en el presente capítulo, se han mantenido los parámetros del algoritmo genético utilizado para la búsqueda conformacional en sus valores de default (150 para el tamaño inicial de la población (population size), 2.5 x 10^6 como el número máximo de evaluaciones de energía y 2.7x10^4 como el número máximo de generaciones). Para cada par proteína-ligando, se creó una grilla utilizando el programa Autogrid de manera tal que el tamaño de la misma incluya todo CBS. Para este propósito, el centro de la grilla ubicó en el centro geométrico del CBS, calculado como la coordenada promedio de todos los átomos distintos de hidrógeno de los residuos correspondientes al CBS. Estos residuos son todos los aminoácidos de la proteína tal que se encuentren a menos de 5 Å del algún átomo distinto de hidrógeno del ligando en la correspondiente estructura de referencia proteína-ligando. Por lo tanto, el tamaño de la grilla se extiende entre 20 (para el caso de proteínas que unen mono- y disacáridos) y 25 Å (para aquellas proteínas que unen tri y tetrasacáridos) sobre cada dirección desde el centro de la grilla. Para cada estructura, se realizaron 100 corridas de docking y los resultados fueron clusterizados de acuerdo a un valor de corte de 2 Å con respecto a un valor de RMSD sobre los átomos

distintos de hidrógenos de los ligandos. Finalmente, fue necesario realizar simulaciones de dinámica molecular sobre la estructura libre de ligandos de CD44 (PDB ID 2JCP)[22] y de la proteína surfactante, SP-D, (PDB ID 2GUU),[23] utilizando el mismo protocolo de simulación que fuera previamente descripto en la sección Metodología Específica del capítulo 3

Referencias

[1] D.S. Goodsell, G.M. Morris, and A.J. Olson. Automated docking of flexible ligands: Applications of autodock. *Journal of Molecular Recognition*, 9(1):1–5, 1996. cited By (since 1996)543.

[2] G.M. Morris, D.S. Goodsell, R. Huey, and A.J. Olson. Distributed automated docking of flexible ligands to proteins: Parallel applications of autodock 2.4. *Journal of Computer-Aided Molecular Design*, 10(4):293–304, 1996. cited By (since 1996)603.

[3] Z. Li and T. Lazaridis. Thermodynamic contributions of the ordered water molecule in hiv-1 protease. *Journal of the American Chemical Society*, 125(22):6636–6637, 2003. cited By (since 1996)54.

[4] M. Agostillo, C. Jene, T. Boyle, P.A. Ramsland, and E. Yuriev. Molecular docking of carbohydrate ligands to antibodies: Structural validation against crystal structures. *Journal of Chemical Information and Modeling*, 49(12):2749–2760, 2009. cited By (since 1996)27.

[5] P. Englebienne and N. Moitessier. Docking ligands into flexible and solvated macromolecules. 5. force-field-based prediction of binding affinities of ligands to proteins. *Journal of Chemical Information and Modeling*, 49(11):2564–2571, 2009. cited By (since 1996)9.

[6] E. Feliu and B. Oliva. How different from random are docking predictions when ranked by scoring functions? *Proteins: Structure, Function and Bioinformatics*, 78(16):3376–3385, 2010. cited By (since 1996)8.

[7] A. Kerzmann, J. Fuhrmann, O. Kohlbacher, and D. Neumann. Balldock/slick: A new method for protein-carbohydrate docking. *Journal of Chemical Information and Modeling*, 48(8):1616–1625, 2008. cited By (since 1996)15.

[8] A. Kerzmann, D. Neumann, and O. Kohlbacher. Slick - scoring and energy functions for protein - carbohydrate interactions. *Journal of Chemical Information and Modeling*, 46(4):1635–1642, 2006. cited By (since 1996)14.

[9] A.R. Leach, B.K. Shoichet, and C.E. Peishoff. Prediction of protein-ligand interactions. docking and scoring: Successes and gaps. *Journal of Medicinal Chemistry*, 49(20):5851–5855, 2006. cited By (since 1996)268.

[10] J. Seco, F.J. Luque, and X. Barril. Binding site detection and druggability index from first principles. *Journal of Medicinal Chemistry*, 52(8):2363–2371, 2009. cited By (since 1996)54.

[11] R.D. Taylor, P.J. Jewsbury, and J.W. Essex. A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, 16(3):151–166, 2002. cited By (since 1996)320.

[12] S.K. Mishra, J. Adam, M. Wimmerová, and J. Koča. In silico mutagenesis and docking study of ralstonia solanacearum rsl lectin: Performance of docking software to predict saccharide binding. *Journal of Chemical Information and Modeling*, 52(5):1250–1261, 2012. cited By (since 1996)3.

[13] D.T. Moustakas, P.T. Lang, S. Pegg, E. Pettersen, I.D. Kuntz, N. Brooijmans, and R.C. Rizzo. Development and validation of a modular, extensible docking program: Dock 5. *Journal of Computer-Aided Molecular Design*, 20(10-11):601–619, 2006. cited By (since 1996)163.

[14] O. Trott and A.J. Olson. Software news and update autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. cited By (since 1996)772.

[15] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004. cited By (since 1996)1302.

[16] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998. cited By (since 1996)4332.

[17] R. Huey, G.M. Morris, A.J. Olson, and D.S. Goodsell. Software news and update a semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry*, 28(6):1145–1152, 2007. cited By (since 1996)649.

[18] C.M.A. Guardia, D.F. Gauto, S. Di Lella, G.A. Rabinovich, M.A. Martí, and D.A. Estrin. An integrated computational analysis of the structure, dynamics, and ligand binding interactions of the human galectin network. *Journal of Chemical Information and Modeling*, 51(8):1918–1930, 2011. cited By (since 1996)5.

[19] M. Agostino, M.S. Sandrin, P.E. Thompson, E. Yuriev, and P.A. Ramsland. Identification of preferred carbohydrate binding modes in xenoreactive antibodies by combining conformational filters and binding site maps. *Glycobiology*, 20(6):724–735, 2010. cited By (since 1996)11.

[20] M. Agostino, E. Yuriev, and P.A. Ramsland. A computational approach for exploring carbohydrate recognition by lectins in innate immunity. *Frontiers in Immunology*, 2(JUN), 2011. cited By (since 1996)4.

[21] G.M. Morris, H. Ruth, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. Software news and updates autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009. cited By (since 1996)776.

[22] S. Banerji, A.J. Wright, M. Noble, D.J. Mahoney, I.D. Campbell, A.J. Day, and D.G. Jackson. Structures of the cd44-hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nature Structural and Molecular Biology*, 14(3):234–239, 2007. cited By (since 1996)75.

[23] E. Crouch, B. McDonald, K. Smith, T. Cafarella, B. Seaton, and J. Head. Contributions of phenylalanine 335 to ligand recognition by human surfactant protein d: Ring interactions with sp-d ligands. *Journal of Biological Chemistry*, 281(26):18008–18014, 2006. cited By (since 1996)19.

Capítulo 6
Discusión General y Conclusión

Como ya se ha planteado en la introducción de esta tesis, las asociaciones no covalentes entre biomoléculas (en particular entre proteínas y moléculas de bajo peso molecular, que llamamos ligandos), tienen un fuerte impacto en numerosos procesos biológicos como ser la acción de hormonas, el control de la transcripción del DNA, el reconocimiento de antígenos por parte del sistema inmune, la catálisis de reacciones químicas por parte de enzimas y la acción de la gran mayoría de los compuestos terapéuticos o drogas. En general, el principio que prima la interacción entre proteínas y carbohidratos contempla por un lado el efecto cooperativo que se produce entre las interacciones de tipo puente hidrógeno entre ambas biomoléculas, y por otro lado las interacciones hidrofóbicas, que en muchos casos vienen dadas por las interacciones π/C-H entre la cara no polar de los carbohidratos (cara-B), y el grupo aromático de algún residuo proteico (por ejemplo, el triptófano presente en el CBS en Galectinas). No obstante, la contribución efectiva de los puentes de hidrógeno a la energía libre de formación del complejo proteína-carbohidrato, no depende únicamente de su interacción intrínseca, sino también de la energía libre asociada a la deshidratación de las superficies de contacto y el concomitante desplazamiento de moléculas de agua desde la superficie de interacción hacia el seno del solvente.

De este modo, es el balance neto de estos procesos, lo que determinará si la formación de un puente de hidrógeno contribuirá de manera significativa a la energía libre de unión. Dada la naturaleza polianfifílica de los carbohidratos y las superficie proteica que los reconoce, el efecto del solvente está más acentuado en este tipo de sistemas. En este contexto, comprender la estructura y dinámica de las moléculas de agua en torno a la superficie proteica es un desafío que puede ser abordado desde distintas perspectivas. Desde el punto de vista experimental, recientemente Saraboji et al. han planteado experimentos combinados de cristalografía y RMN de alta resolución, que permiten discernir la posición y la dinámica de las moléculas en el CBS de Galectina-3, para luego relacionar la posición de las mismas con la estructura del complejo Gal3-lactosa, habiendo observado que el CBS de Gal-3 está preorganizado por las moléculas de agua que luego serán desplazadas por los OHs del carbohidrato cuando se forma el complejo.[1] Por otro lado, Howard et al., a partir

de experimentos de difracción de neutrones en conjunto con experimentos de difracción de rayos-X, reportaron la estructura detallada de la proteína anticongelante de tipo III junto con su estructura de solvatación.[2] Como ya se ha visto en el contexto de está tesis, las simulaciones de dinámica molecular proveen el detalle concerniente al movimiento de las partículas en función del tiempo y por consiguiente, son utilizadas para explorar el espacio conformacional del sistema en estudio, obteniéndose una descripción microscópica del mismo en su estado de equilibrio, incluyendo la observación de propiedades vinculadas a la estructura y la dinámica de los sistemas, así como la posible derivación de parámetros termodinámicos. Esta información, puede ser utilizada para responder interrogantes acerca de las propiedades del sistema en estudio.

En la presente tesis, se ha planteado la hipótesis general de que las interacciones que las moléculas de agua establecen con la superficie del CBS emulan las interacciones de los grupos OHs de los carbohidratos en el complejo correspondiente. Entonces, se determinaron las propiedades termodinámicas de las mismas para predecir qué moléculas de agua tienen mayores chances de ser desplazadas desde el CBS hacia el seno del solvente, y con está información mejorar la predicción de la estructura e interacciones de los complejos proteína-carbohidrato. En el capítulo 3, se puso a prueba la hipótesis de trabajo analizando un conjunto de 5 proteínas, 4 de ellas unen carbohidratos y la ciclofilina A que une al decapéptido cíclico, ciclosporina. Como se ha mostrado, es posible en base al cálculo de las funciones de distribución radial y angular, definir la estructura de solvatación en el CBS, y la caracterización del mismo en función de los WS (que se definen como aquellas regiones del espacio aledañas a la superficie proteica tal que la probabilidad de encontrar moléculas de agua con respecto al seno del solvente fuera mayor a 1). Luego se determinaron las propiedades estructurales y termodinámicas de los WS y se compararon con la posición de los grupo OHs en el complejo. **En base a esto, se pudo comprobar que la probabilidad relativa al seno del solvente de encontrar moléculas de agua en el WS (WFP) y el radio tomado desde el centro del WS tal que incluya al 90 % de las moléculas de agua que incluye el mismo (R_{90}), son los mejores predictores de la posición de los grupos OH del ligando**. Estos parámetros, a diferencia de la energía de interacción agua/proteína, incluyen de manera implícita un componente entrópico.

Está observación nos permite abordar ahora un análisis más profundo sobre cuáles son las contribuciones termodinámicas más relevantes que caracterizan a los WS. Para ello, asumiendo que las moléculas de agua en los WS pueden ser descriptas en base a una distribución de tipo Boltzmann, es posible estimar el cambio en la energía libre de llevar una molécula de agua desde el seno del solvente hacia el WS, como:

$$\Delta G_{WS} = -RTln(WFP) \tag{6.1}$$

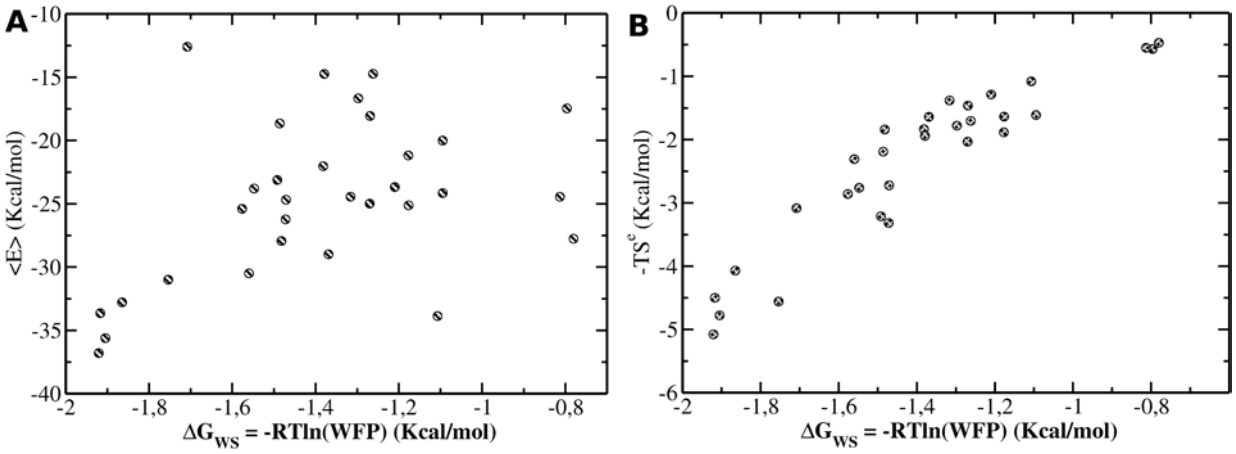

Figura 6.1. Gráficos de correlación entre A) la energía de interacción media de las moléculas de agua con su entorno <E>vs. -RTln(WFP) y B) el término que incluye la entropía de exceso parcial, -TS^e vs. -RTln(WFP) para los WS que son reemplazados por grupos OHs de los carbohidratos (R_{min} <1.5 Å) para un set de más de 15 lectinas diferentes que unen una variedad de ligandos sacarídicos (desde mono hasta tetra-sacáridos)

y correlacionar el valor obtenido con los valores de entropía de exceso y de energía de interacción de las moléculas de agua en el WS calculadas en base a la teoría del fluido inhomogéneo tal como se detalla en el capítulo de métodos (capitulo 2). En la figura 6.1 A y B, se muestra la relación entre la estimación de $< E_T >$ vs. ΔG_{WS} y $-TS_{WS}$ vs. ΔG_{WS}, respectivamente. **De la misma, es posible inferir que el término WFP correlaciona mucho mejor con el término entrópico que con el de energía de interacción (adicionalmente, el calculo del coeficiente de correlación de Pearson, R, es de 0.6 y de 0.9 para las figuras 6.1 A y B respectivamente), lo cuál sugiere que la mayor contribución a WFP es entrópica**. Esto apoya la hipótesis planteada por numerosos autores de que las aguas que se encuentran mas retenidas por la superficie proteica son las que mayores chances tienen de ser desplazadas, debido la ganancia entrópica dada por desplazamiento de las mismas hacia el seno del solvente. Dicho de otro modo, hemos visto en el capitulo 3 que a mayor WFP y menor R_{90}, las chances que tiene un agua de ser desplazada por un OH del ligando son mayores. Esto podría explicarse entonces en base a que la penalidad entrópica impuesta por alojar una molécula de agua en un WS (la cual es mayor cuanto mayor es el valor de WFP y menor el de R_{90}), y por lo tanto mayor es la ganancia entrópica al ser liberada cuando es reemplazada por el grupo OH del ligando. El hecho de que estos dos parámetros tengan una buena correlación con el R_{min} (que mide la distancia del centro del WS al átomo más cercano del ligando distinto de hidrógeno en el complejo), permite establecer un criterio de qué WS tendrán la mayores chances de ser reemplazados por los OHs del ligando.

Por otro lado, la correlación con los parámetros termodinámicos determinados, com-

parados contra los realizados a partir de una teoría mas robusta como es la teoría del fluido inhomogéneo (IFST), permite utilizarlos con mayor grado de confianza y con la ventaja de que el calculo de los mismos es computacionalmente menos costoso en comparación al cálculo de los parámetros termodinámicos a partir de IFST, que es más demandante desde el punto de vista computacional. Los algoritmos que computan las propiedades de los WS, escalan mucho más lentamente, O(N^3) vs. O(N^2) (con N el número de átomos de oxígeno de las moléculas de agua) para el IFST y el cálculo de WFP, respectivamente.

Dado que la mayor parte de los carbohidratos difieren en la configuración relativa de uno o dos de sus grupos hidroxilo, la selectividad epimérica tiene importantes consecuencias en los procesos de reconocimiento biológico a nivel molecular. Es así como el establecimiento de una red cooperativa entre grupos donores/aceptores de puentes de hidrógeno además de la complementariedad topológica con la proteína terminara determinando la unión, modulada por la orientación axial/ecuatorial de los grupos OH. **No obstante, como se ha visto en el capítulo 4, para delinear los detalles de la alta precisión del proceso de reconocimiento, es necesario incluir los efectos explícitos del solvente**. Por un lado hemos visto que ciertas moléculas de agua aumentan la complementariedad entre las superficies de ambas biomoléculas incrementando de este modo la energía de interacción directa entre las mismas, mientras que otras son desplazadas hacia el seno del solvente pudiendo influir favorablemente a la energía libre de unión. Estos dos efectos, están principalmente determinados por la orientación espacial que las moléculas de aguas adoptan en la proteína libre, lo que guarda una relación directa con la orientación espacial de los grupo OH en el ligando.

La predicción exacta de los modos de unión entre un ligando y el blanco macromolecular (docking) es de central importancia en el diseño de drogas basado en la estructura y en las últimas décadas los métodos computacionales de docking han sido ampliamente utilizados para este propósito. Los resultados de los capítulos 3 y 4 muestran que si se tiene en cuenta el desplazamiento de las moléculas de agua asociadas a superficies hidrofílicas en el proceso de unión, esto daría un modelo más realista de la estructura del complejo y de la selectividad entre ligandos. Esto nos lleva a plantearnos la inclusión de este proceso en los esquema de docking convencionales, los cuales modelan la desolvatación de manera aproximada en base a cálculos que tienen en cuenta la diferencia en la superficie expuesta a solvente de las moléculas interactuantes. Si bien se ha visto que para ligandos tipo drogas (es decir, ligandos con un alto carácter hidrofóbico) esto es una buena aproximación, para ligandos y superficies proteicas polianfifílicas se requiere tener en cuenta un modelo mas realista de solvente. **Los resultados del capítulo 5 muestran que la modificación de la energía de libre de interacción que computa el programa de docking conlleva a una mejora significativa en el carácter predictivo del mismo tanto en términos de**

exactitud, es decir la capacidad de predecir el complejo correcto, y precisión, es decir la capacidad de discernir la estructura del complejo coincidente con la estructura de referencia cristalográfica de los números resultados falsos positivos que arroja el programa. Por otro lado, es relevante resaltar que en la mayor parte de los casos se ha observado que el método convencional de docking (CADM) es capaz de encontrar el complejo correcto (es decir, que el algoritmo genético muestrea el espacio configuracional del complejo de manera tal que uno de los complejos propuestos coincide con el complejo correcto). No obstante, al momento de calcular el ΔGu, no lo considera como un complejo de buena interacción penalizándolo en el ranking propuesto. Con el método modificado de docking (WSBDM), esta falla es corregida y ahora el complejo correcto es reportado en la mayor parte de los casos como el complejo de mayor afinidad y por lo tanto, premiado en el ranking de todos los posibles complejos propuestos.

6.0.1. Perspectivas Futuras

En rasgos generales, el hilo conductor de los capítulos de resultados de la presente tesis muestran que la estructura y dinámica de las moléculas de agua en la proteína libre, obtenidas a partir de simulaciones por dinámica molecular, es capaz de mapear las interacciones que luego serán establecidas por los grupos OHs de los carbohidratos en el complejo, posiblemente como reflejo de la similitud entre la interacción de los grupos OH de las moléculas de aguas con la proteína respecto de la de los grupos OH de los azúcares. Por otro lado, existe evidencia experimental (a partir de experimentos de RMN y de cristalografía de rayos-X) que los sitios de unión de ligandos de las proteínas tienden a unir pequeñas moléculas de solventes orgánicos (como por ejemplo, el mapeo de los sitios de unión la lisozima de huevo por pequeñas moléculas como metanol cloruro de metileno, acetonitrilo, acetona, DMSO, isopropanol y urea mediante RMN,[3] y el mapeo de la superficie de unión a ligandos de la elastasa pancreática bobina por medio de la resolución de la estructura cristalográfica con acetonitrilo como cosolvente).[4] Esta observación sugiere que es posible ampliar la hipótesis de trabajo de la presente tesis, planteando una estrategia análoga para el mapeo de distinto tipo de interacciones con pequeñas moléculas. Para ello, a diferencia de lo que se hizo en esta tesis, se pueden simular por dinámica molecular proteínas embebidas en soluciones de cosolvente/agua en donde el soluto debiera elegirse de manera tal que muestree las interacciones en la superficie proteica deseadas. Por ejemplo, si se requiere explorar las porciones hidrofóbicas de la misma, se puede simular a la proteína en soluciones de etanol/agua o bien isopropanol/agua en donde las características de estás moléculas de cosolvente cumplen con ser lo suficientemente pequeñas para asegurar un coeficiente de difusión razonable en la escala de tiempo de la simulación y además, las características anfipáticas de los alcoholes les darían la ventaja de ser parcialmente miscibles

en agua. De este modo, la elección del cosolvente debiera estar basado en las propiedes fisicoquímicas del mismo de manera tal que sea capaz de interaccionar con la superficie proteica poniendo de relieve aquellas regiones que se requieran detectar (polares, cargadas negativa/positiva, no polares, aromáticas, etc). Luego, el análisis de estás dinámicas mixtas podrían analizarse perfectamente con el programa de clusterización presentado en el capítulo de métodos de esta tesis, pero con un sutil modificación para que el análisis se efectué sobre el cosolvente en vez de sobre las moléculas de agua circundantes.

Por otro lado, dado el rol de los carbohidratos en el metabolismo y como piezas estructurales, los mismos ofrecen una alternativa no explotada para el diseño de drogas y por consiguiente abren el camino para plantear nuevas oportunidades terapéuticas. No obstante, la afinidad de los carbohidratos por las lectinas es baja como consecuencia de que el sitio de unión es poco profundo y polar. Esto motiva a diseñar racionalmente de fármacos glicomiméticos, que tengan mayor afinidad y selectividad que los compuestos naturales, lo que debe hacerse en base a un conocimiento detallado de los principios que gobiernan la interacción proteína-carbohidrato a partir de las cuales pueden plantearse diferentes aproximaciones que van desde el diseño de drogas cuya conformación en el estado libre este restringida a la de la forma bioactiva en el complejo (y con eso reducir la penalidad entrópica que infiere la perdida de los grados de libertad conformacionales en la forma unida), hasta elegir la configuración apropiado en lo concerniente a la orientación de los grupos hidroxilo de la droga glicomimética (con el fin de desplazar determinadas moléculas de agua del CBS, como también aprovechar otras moléculas de agua que puenteen la interacción mejorando la complementariedad estructural droga-proteína dentro del contexto y las hipótesis que se han planteado en esta tesis).

6.0.2. Extensión a otros sistemas

Dado que el agua es un componente muy versátil y promiscuo en la naturaleza, el análisis de la estructura del solvente y de las propiedades termodinámicas de las moléculas de agua que forman parte de esta estructura, tiene alcances que van más allá de las lectinas, de modo tal que en los últimos años ha existido un interés creciente en desarrollar descriptores cuantitativos de las propiedades termodinámicas que incluyan los efectos de la hidratación en la unión de ligandos en general.[5–7] Por ejemplo, el estudio de la estructura de solvatación de la hemoglobina truncada de la bacteria *thermobifida fusca* reveló que desplazar moléculas de agua por parte de pequeños ligandos como por ej O_2, CO, etc. tiene un fuerte impacto sobre la cinética de la reacción de unión del CO al hierro del grupo hemo y por ende, se ha visto que el diseño de mutantes que estabilicen menos las moléculas de agua en el sitio activo aumenta la velocidad de reacción.[8] Otro caso en el que se ha aplicado esta estrategia de análisis, es para la proteína transmembrana OmpC que es una

acuoporina de *E. coli* en la cual se ha estudiado la estructura de solvatación del canal transmembrana y donde muestran que la velocidad del pasaje de un antibiótico a través de el mismo depende fuertemente de la facilidad con la se desplazan moléculas de agua dentro del canal y de reemplazar las interacciones agua/proteína por antibiótico /proteína.[9] Finalmente, es posible extender esta estrategia de análisis para estudiar la estructura de solvatación de sistemas materiales como TiO_2 y SiO_2, las cuales suelen mostrar en sus superficies regiones polares expuestas a solvente que forman fuertes interacciones con el agua interfacial, tendiendo a inmovilizar la capa adyacente de solvente (de modo tal que las propiedades físicas del agua interfacial se pueden pensar como una fase de agua liquida condensada sobre la superficie) mientras que las capas sucesivas se asemejan más a las propiedades del agua en el seno del solvente a medida que se aleja de la superficie del sólido.[10]

6.1. Conclusión

En resumen, los tópicos explorados en esta tesis sientan las bases y generan las herramientas para abordar la descripción y el estudio de los parámetros termodinámicos de las moléculas de agua en la superficie proteica relacionadas con el proceso de unión proteína-carbohidrato. Por otro lado, como ha sido ampliamente discutido, la energía libre de unión de carbohidratos a lectinas se encuentra entre 4 y 5 Kcal/mol, principalmente para monosacáridos y disacáridos, no existiendo una tendencia clara que permita decir que los monosacáridos unen más débilmente que los disacáridos. En este sentido, es esperable que una descripción termodinámica rigurosa de los cambios de entorno del solvente de las biomoléculas libres hacia la formación del complejo, pueda contribuir de manera significativa a la mejora de los métodos actuales de cálculo de energía libre proteína-carbohidrato.

Finalmente, una aspecto de carácter general, aunque no menos importante que creo relevante resaltar al final de esta tesis, es la importancia del enfoque multidisciplinario como estrategia de investigación, en donde las diversas disciplinas, (desde las ciencias físicas, biológicas y químicas hasta ciencias de la computación), se concatenan para alcanzar un determinado interrogante generando un efecto sinérgico en el que el todo es más que la suma de las partes, sin el cual los aspectos discutidos y abordados durante estos años me hubieran resultado inalcanzables.

Referencias

[1] K. Saraboji, M. Hakansson, S. Genheden, C. Diehl, J. Qvist, U. Weininger, U.J. Nilsson, H. Leffler, U. Ryde, M. Akke, and D.T. Logan. The carbohydrate-binding site in galectin-3 is preorganized to recognize a sugarlike framework of oxygens: Ultra-high-resolution structures and water dynamics. *Biochemistry*, 51(1):296–306, 2012. cited By (since 1996)9.

[2] E.I. Howard, M.P. Blakeley, M. Haertlein, I.P.-. Haertlein, A. Mitschler, S.J. Fisher, A.C.-. Siah, A.G. Salvay, A. Popov, C.M.-. Dieckmann, T. Petrova, and A. Podjarny. Neutron structure of type-iii antifreeze protein allows the reconstruction of afp-ice interface. *Journal of Molecular Recognition*, 24(4):724–732, 2011. cited By (since 1996)12.

[3] E. Liepinsh and G. Otting. Organic solvents identify specific ligand binding sites on protein surfaces. *Nature Biotechnology*, 15(3):264–268, 1997. cited By (since 1996)74.

[4] K.N. Allen, C.R. Bellamacina, X. Ding, C.J. Jeffery, C. Mattos, G.A. Petsko, and D. Ringe. An experimental approach to mapping the binding surfaces of crystalline proteins. *Journal of Physical Chemistry*, 100(7):2605–2611, 1996. cited By (since 1996)109.

[5] J. Michel, J. Tirado-Rives, and W.L. Jorgensen. Prediction of the water content in protein binding sites. *Journal of Physical Chemistry B*, 113(40):13337–13346, 2009. cited By (since 1996)42.

[6] P. Setny, R. Baron, and J.A. McCammon. How can hydrophobic association be enthalpy driven? *Journal of Chemical Theory and Computation*, 6(9):2866–2871, 2010. cited By (since 1996)50.

[7] R. Baron, P. Setny, and J. Andrew McCammon. Water in cavity-ligand recognition. *Journal of the American Chemical Society*, 132(34):12091–12097, 2010. cited By (since 1996)66.

[8] J. Bustamante, S. Abbruzzetti, A. Marcelli, D. Gauto, L. Boechi, A. Bonamore, A. Boffi, S. Bruno, A. Feis, P. Foggi, D. Estrin, and C. Viappiani. Ligand uptake modulation by internal water molecules and hydrophobic cavities in proteins. *Submitted*, 2013.

[9] Q.-T. Tran, S. Williams, R. Farid, G. Erdemli, and R. Pearlstein. The translocation kinetics of antibiotics through porin ompc: Insights from structure-based solvation mapping using watermap. *Proteins: Structure, Function and Bioinformatics*, 81(2):291–299, 2013. cited By (since 1996)1.

[10] E.G. Solveyra, E. De La Llave, V. Molinero, G.J.A.A. Soler-Illia, and D.A. Scherlis. Structure, dynamics, and phase behavior of water in tio2 nanopores. *Journal of Physical Chemistry C*, 117(7):3330–3342, 2013. cited By (since 1996)2.

Capítulo 7
Publicaciones

Trabajos publicados en el contexto de esta tesis

- *S*olvent structure improves docking prediction in lectin-carbohydrate complexes. **Gauto, D. F.**; Petruk, A. A.; Modenutti, C.P., Blanco, J.I.; Di Lella, S.; Marti, M. A. Glycobiology, 23 (2), pp. 241-258, 2013.
- *D*etemining free energy of protein-ligand binding and association/dissociation processes using computer simulations. **Gauto, D.F.**, Modenutti C., Dumas, V.G., Alvarez, L., Bustamante, J.P., Turjankski, A.G., Marti, M.A., World Research Journal of Peptide and Protein, 1 (1), pp. 21-32, 2012
- *S*tructural basis for ligand recognition in a mushroom lectin: Solvent structure as specifity predictor. **Gauto, D. F.**; Di Lella, S.; Estrin, D.A.; Monaco, H.L.; Marti, M.A., Carbohydrate Research, 346(7), pp. 939-948, 2011.
- *C*arbohydrate-binding proteins: Dissecting ligand structures through solvent enviroment occupancy. **Gauto, D. F.**; Di Lella, S.; Guardia, C.M.A.; Estrin, D.A.; Marti, M.A., Journal of Physical Chemistry B, 113(25), pp. 8717-8724, 2009.

Otras publicaciones

- *A*n integrated computational analysis of the structure, dynamics and ligand binding interactions of the human galectin network. Guardia, C.M.A., **Gauto, D. F.**, Di Lella, S., Rabinovich, G. A., Marti, M.A., Estrin, D.A. Journal of Chemical Information and Modeling, 51 (8), pp. 1918-1930, 2011.

Agradecimientos

El fin de esta tesis refleja el trabajo de muchos años , desde el primer día que puse un pie en la Universidad hasta el momento en el que escribo esta líneas. En en lo personal, creo que necesitaría decenas de páginas para poder expresar mi profundo agradecimiento,

A la Universidad pública y a toda la gente que a través de su esfuerzo colectivo hace posible que la educación libre y gratuita sea un sueño hecho realidad.
Al CONICET, que me brindo las becas que hicieron posible mi sustento durante estos años.
A la FCEN, INQUIMAE Y Al DQIAyQF por haberme brindado una formación de excelencia, el marco institucional que me permitió hacer mi doctorado y por el privilegio de haber podido dar clases durante tantos años.

Quisiera expresar mi agradecimiento a mi Director de tesis, Marcelo Martí, por haber confiado en mí desde el primer momento que empecé a trabajar con él, aún como estudiante de grado y por estar siempre dispuesto a buscar soluciones a todas las inquietudes que le planteé a lo largo de estos años, desde las profesionales hasta las personales. Demás está decir que Marcelo es un gran científico, pero los que lo conocemos día a día sabemos que primero es una gran persona con buenos valores, lo que hace que siempre sea un gusto trabajar con él.

A mi amigo Charly Guardia, que ya es familia; a aquellos que siempre me brindaron su amistad y a todo el GMM, por el buen clima de trabajo y la buena predisposición constante.

A toda mi familia, y en especial a Chichi, por ser un ejemplo de vida, pero por sobre todo a Cynthia por su apoyo incondicional y por todo el amor que me da día a día y a mi hijo Bautista por demostrarme que cada segundo de la vida es hermoso. Sin ustedes, nada de todo esto tendría sentido.