DESARROLLO DE METODOLOGÍAS PARA EL ESTUDIO DE MODELOS SIMPLIFICADOS DE ADN-PROTEÍNA

Tesis de Maestría en Bioinformática

18 de Junio 2015

Lic. Astrid Brandner

Orientador: Dr. Sergio Pantano

Co-orientador: Dr. Francisco Melo

Trabajo realizado en el Grupo de Simulaciones Biomoleculares

Institut Pasteur de Montevideo

TABLA DE CONTENIDO

RESUMEN

Las interacciones ADN-proteína juegan un papel vital en innumerables sistemas biológicos como el empaquetamiento de los cromosomas, la reparación del ADN, transcripción o represión génica. El estudio estructural de este tipo de sistemas ha sido abordado utilizando una gran variedad de métodos teóricos y de modelización. Sin embargo, la caracterización de las interacciones que regulan la afinidad y especificidad del reconocimiento ADN-proteína es todavía incompleta, en parte debido a que las interacciones relevantes se extienden en un amplio rango de valores en las escalas temporal y espacial.

En el grupo de Simulaciones Biomoleculares del Institut Pasteur de Montevideo se ha desarrollado recientemente un conjunto de modelos moleculares simplificados (denominado SIRAH) en los que centros efectivos de interacción representan grupos de átomos. La disminución del número de componentes de los sistemas moleculares permite una drástica reducción de los tiempos de cálculo manteniendo la esencia de las interacciones fisicoquímicas que rigen los sistemas moleculares. Los potenciales que determinan las interacciones entre centros efectivos son descriptos por una función Hamiltoniana idéntica a la utilizada en la gran mayoría de paquetes de simulación de uso libre, lo que facilita su implementación. Este tipo de representaciones simplificadas, también llamadas de grano grueso o *coarse-grained* (CG), permiten el estudio *in silico* de la evolución temporal de agregados macromoleculares en tiempos y tamaños comparables a los biológicamente relevantes.

En este proyecto de maestría se propone la caracterización, desarrollo y aplicación de este tipo de técnicas para describir el proceso de reconocimiento de ADN y proteínas utilizando los parámetros del campo de fuerza SIRAH. Para esto fue necesario derivar parámetros usando la información estructural de complejos ADN-Proteína reportados en la base de datos Protein Data Bank (PDB http://www.pdb.org). Esto se realizó mediante base de datos PDIdb (http://melolab.org/PDIdb) que contiene un conjunto curado de estructuras de complejos ADN-proteína y una clasificación funcional basada en datos estadísticos de cada átomo presente en la interfase entre macromoléculas. Estudios realizados sobre una serie de complejos revelaron que a pesar de una pérdida de algunos contactos nativos existe una alta conservación de motivos de estructura secundaria así como a nivel de RMSD de regiones estructuradas de la proteína y de los fosfatos del ADN. El presente trabajo de tesis se centra en resultados y desarrollos no publicados aun. Para una descripción detallada de resultados publicados se remite al lector al Anexo II.

TRABAJOS DERIVADOS EN EL MARCO DE LA TESIS

Presentaciones:

- “Assessing DNA-polylysine and DNA-polyarginine interactions with coarse grain molecular dynamics ” (poster), PASI: Molecular-Based Multiscale Modeling and Simulation, Institut Pasteur de Montevideo, 8 de Setiembre de 2012
- “Dual resolution MD simulations: Fine-grain solute/Coarse-Grain aqueous solvent” (poster y presentaciones orales cortas), Multiscale Modeling methods for applications in material science, Jülich, 17 de Setiembre de 2013 y Espresso Summer School, Universität Stuttgart, 8 de Octubre de 2013
- "Towards the rational design of protein vectors in gene therapy" (poster), Café Bioinformático, Jornadas organizadas por PEDECIBA Bioinformática, Facultad de Química, UdelaR, 05 de Junio de 2014
- “SIRAH proteins: knowledge-based coarse grain force field for proteins” (poster), EMBO Practical Course in Biomolecular Simulations, Institut Pasteur, Paris, 20-27 de Julio de 2014
- "Brief Introduction to SIRAH coarse-grained force field for proteins" (presentación oral), Theoretical and Computational Biophysics Department, Max Planck Institute for Biophysical Chemistry, Göttingen, Alemania, 01 de Agosto de 2014

Publicaciones:

Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level. Pablo D. Dans, Leonardo Darré, Matías R. Machado, Ari Zeida, Astrid F. Brandner, Sergio Pantano. Lect. Notes in Comp. Science , 2013. DOI:10.1007/978-3-319-02624-4_7

SIRAH: a structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. Leonardo Darré, Matías R. Machado, Astrid F. Brandner, Humberto C. González, Sebastián Ferreira, and Sergio Pantano. J. Chem. Theory Comput., 2015. DOI:10.1021/ct5007746

Registro de propiedad intelectual:

The SIRAH force field 2014. Autores: Sergio Pantano, Matías Machado, Astrid Brandner, Humberto González, Leonardo Darré, Pablo Dans, Ari Zeida. Noviembre 2014.

Los resultados de esta tesis forman parte de un artículo en preparación.

OBJETIVOS

Objetivo general

Desarrollar un conjunto de parámetros para interacciones CG ADN-Proteína de validez general. Dichos parámetros serán derivados de información estadística recopilada del PDB.

Objetivos específicos

- Verificación y eventual optimización del campo de fuerzas CG para proteína.
- Ajuste de parámetros específicos de interacción ADN-Proteína.
- Validación de los potenciales de interacción derivados para complejos ADN-Proteína representativos.

TERMINOLOGÍA ESPECÍFICA UTILIZADA

Los aminoácidos en coarse-grained en este trabajo se denominan con el código habitual de una letra antecedido por la letra s minúscula. Ej: alanina coarse-grained se denomina sA. En caso de analizarse un resultado atomístico, los nombres de los aminoácidos se presentan con su código de 3 letras para facilitar la comprensión.

Los nombres de los átomos coarse-grained siempre siguen la nomenclatura presentada en la tabla 2.2 (pág. 25). Por ejemplo, los átomos coarse-grained llamados GC corresponden a Cα y los PX a P, representando al fosfato del ADN.

atomtypes: tipos de átomos en un campo de fuerza que poseen parámetros distintivos

Beads: Pseudoátomos del modelo de grano grueso. También pueden aparecer como "átomos CG"

CG: coarse-grained

FG: modelo donde se representan con detalle atomísitco a las moéculas (fine-grained)

IHF: Factor de integración al hospedero

PDB: Protein Data Bank (http://www.pdb.org)

PDB ID: Código de estructura PDB

PDIdb: Protein Interface database

pseudoátomos: átomos (partículas) del campo de fuerza de grano grueso.

RMSD: Desviación cuadrática media (Root Mean Square Deviaton)

TBP: Proteína de unión a caja TATA

WT4: modelo de agua coarse-grained explícita presente en el campo de fuerza SIRAH

I INTRODUCCIÓN

1.1: Introducción general

Múltiples procesos esenciales en la vida celular involucran las interacciones de proteínas con ADN. Este es el caso desde el empaquetamiento de la información nuclear en cromosomas y cromatina en el caso de organismos eucariotas, a numerosos procesos de transcripción, duplicación, embriogénesis y expresión génica en general. Debido a la importancia de estos sistemas, se han utilizado una gran variedad de métodos teóricos y de modelización para estudiarlos con detalle.

Sin embargo, la caracterización a nivel microscópico de las interacciones que regulan la afinidad y especificidad del reconocimiento ADN-proteína se encuentra todavía incompleta, en parte debido a que las interacciones relevantes se extienden en un amplio rango de valores en las escalas temporal y espacial. En este contexto, el modelado computacional combinado con modernas técnicas de simulación molecular se presentan como una alternativa sólida para subsanar dificultades experimentales y permitiendo la integración de datos de distinta naturaleza para obtener información estructural. En particular, las técnicas de dinámica molecular permiten un seguimiento espacial y temporal de estos complejos, aunque encuentran serias dificultades en cuanto al costo computacional de dichas simulaciones. De este modo, el poder de cálculo disponible limita la extensión de los sistemas capaces de ser modelados a tamaños y tiempos que son, en general, demasiado pequeños para ser comparados con escalas biológicamente relevantes.

Con el objetivo de superar estos problemas, se han propuesto distintas aproximaciones que reducen la complejidad y grados de libertad de las macromoléculas biológicas, disminuyendo sensiblemente el costo computacional. En este trabajo se verá el enfoque de métodos simplificados utilizado que corresponde a un modelo de grano grueso (abreviado como CG del inglés *coarse-grained*).

1.2: Generalidades estructurales del ADN

Cuando comúnmente se piensa en la secuencia de ADN se tiende a considerarla únicamente en análisis unidimensionales, omitiendo el efecto que tiene en la estructura. Dado que en este trabajo el enfoque está sobre la estructura de complejos ADN-proteína es de particular interés resumir brevemente algunas propiedades estructurales genéricas del ADN así como también secuencia-dependientes.

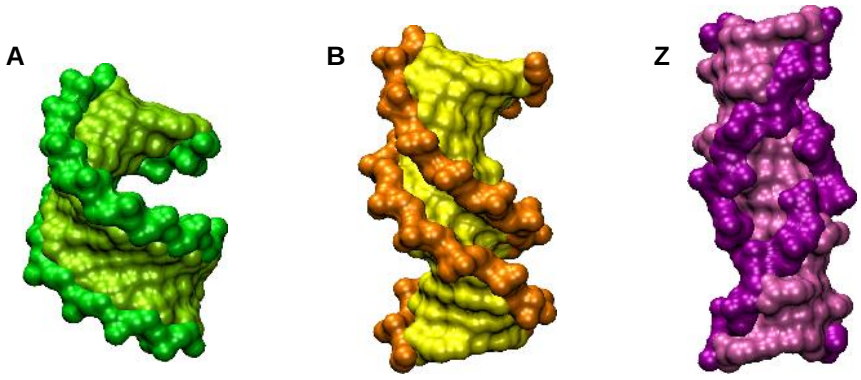
El ADN canónico de doble hebra está estructuralmente conformado por dos moléculas complementarias en forma de doble hélice que establecen un surco menor y un surco mayor

con un esqueleto de fosfatos.

Además de la estructura canónica en forma B, también son conocidas las conformaciones A y Z, así como eventos de distorsiones locales tales como *bend*s o *kink*s. En la tabla 1.1 se resumen las principales propiedades de las estructuras más estudiadas de ADN doble hebra.

Tabla 1.1: Características morfológicas de confórmeros de ADN. Los valores medidos para surcos mayor y menor están definidos en distancias entre fosfatos de ambas hebras. Entre paréntesis se reportan los valores en presencia de *narrowing* para el surco menor. En la parte inferior se muestran las estructuras de un ADN en forma A (PDB ID: 4IZQ), en forma B (PDB ID: 1BNA) y en forma Z (PDB ID: 4OCB).

Conformación del ADN	Ancho aproximado del surco mayor (Å)	Ancho aproximado del surco menor (Å)	Pares de base por vuelta	Diámetro (Å)	Sentido de vuelta	Característica
A	10	20	11	25.5	dextrógiro	Presente en muestras deshidratadas de ADN
B	22	14 (9-10)	10,4	23.7	dextrógiro	ADN canónico
Z	Surco mayor no diferenciable	Surco menor no diferenciable	12	18.4	levógiro	Usualmente repeticiones de dinucleótidos CG



En algunos casos secuencias ricas en AT pueden producir un efecto de mayor curvatura con respecto a la estructura B canónica. El eje central del ADN cuando no se encuentra en su forma lineal y posee una curvatura apreciable determina una existencia de un evento denominado en inglés como *bending.* En estos casos se observa también generalmente un mayor estrechamiento a nivel del surco menor, modificando consecuentemente el potencial electrostático en esa región, favoreciendo así una mayor interacción de residuos catiónicos o con potenciales positivos en la región así como también iones. Por ejemplo estructuras como el nucleosoma o la proteína E2 del virus de papiloma humano tienen una interfase con el ADN que presenta un *bending* marcado, donde principalmente se sitúan los aminoácidos cargados como arginina y lisina a lo largo del surco menor.

Otro tipo de modificación estructural de un ADN canónico lineal es la presencia de los *kinks* (del inglés “quiebres”). Se dice que la estructura de ADN posee un *kink* si existe una interrupción abrupta de una hélice que de otro modo sería lineal. Es una distorsión local a diferencia del *bending* que corresponde a una serie de pequeñas variaciones acumuladas a lo largo de una región del ADN. Por ejemplo el represor lac genera un *kink* en el ADN del operon Lac en la región correspondiente entre ambos dominios de unión al ADN.

En términos generales, las regiones con *kinks* se pueden encontrar formando complejos con proteínas que pueden estabilizar las bases que hayan quedado desapareadas a causa de la modificación estructural, por lo que este tipo de contacto se da a nivel del surco mayor. También puede suceder que regiones con *bending* puedan ayudar a estabilizar las zonas con *kinks* y sean complementarias al evento de quiebre, facilitando una mayor deformación a nivel general del ADN.

Otro ejemplo de cambios estructurales a nivel del ADN son los eventos de estrechamiento del surco menor (denominados *narrowing*). Estas regiones pueden estar presentes en secuencias consecutivas de adeninas que de una manera dinámica a través de presencia de iones o agua en el surco menor presentan esta estructura. Dichas interacciones pueden explicarse puesto que la existencia de *narrowing* genera un aumento del potencial electrostático local, que puede ser estabilizado bajo presencia de iones o aminoácidos catiónicos.

1.3: Motivos estructurales en proteínas de unión al ADN

1.3.1: Mayormente alfa hélices

a) HTH (hélice-giro-hélice)

Este motivo se observa frecuentemente en varias proteínas de unión a ADN. La hélice que reconoce el ADN se une a través del surco mayor mediante enlaces de hidrógeno e interacciones hidrofóbicas mientras que la otra hélice no presenta un rol importante en el reconocimiento del ADN y simplemente estabiliza la interacción. Existen algunas excepciones donde la hélice de reconocimiento se une a lo largo del surco menor. Dentro de estos motivos estructurales existen también los denominados “winged helix-turn helix” (múltiples factores de transcripción de resistencia a antibióticos, de tipo MarR, dominios Ets, entre otros), caracterizados por tener anexado al motivo HTH una corta hebra beta que generalmente está en contacto con el surco menor estabilizando de manera complementaria al complejo.

b) Hélice-loop-hélice y motivos de tipo leucine zipper

Este motivo consiste en una alfa hélice corta seguida de un loop que la comunica con otra alfa hélice mayor. En general la segunda alfa hélice suele dimerizar con otra alfa hélice de igual formando un homodímero. La generación del dímero facilita la estabilización del complejo

ya que ambas hélices largas de cada uno de los monómeros interactúa de manera específica reconociendo su secuencia blanco en el surco mayor del ADN al que se unen.

1.3.2: Mayormente hojas beta

a) TBPs (TATA-box binding proteins)

Estas proteínas reconocen al ADN a través de una hoja beta que interactúa a lo largo del surco menor del ADN. Este motivo es particularmente interesante porque esta interacción solo es posible a causa un alto grado de desviación del ADN de su conformación B hacia un estado de curvatura pronunciado (*bending*) que permite la complementariedad de estructuras y accesibilidad hacia la bases desde el surco menor.

b) Proteinas de tipo Inmunoglobulina

El *folding* general de la proteína es de tipo beta-sandwich, donde el reconocimiento del ADN se da a través de los loops entre las hebras. Ejemplo de este tipo de proteínas pueden ser factores de transcripción de tipo p53.

1.3.3: Proteínas alfa/beta

a) Dedos de zinc (Zinc fingers)

Este tipo de proteínas están compuestas por un dominio de aproximadamente 30 aminoácidos que une al ADN. Éste a su vez contiene una corta hélice alfa, una hoja beta y un ion de Zn^{2+} que se encuentra coordinado por cisteínas y/o histidinas. La alfa hélice es el motivo de reconocimiento del ADN que permite establecer contactos secuencia-específicos con el ADN.

b) RHH (ribbon-helix-ribbon)

Este motivo se caracteriza por tener estructuras de tipo beta-hairpin (2 hebras betas antiparalelas) seguidos por dos hélices. Se encuentra presente en varios factores de transcripción de bacterias (por ejemplo en el represor de Met MetJ). El reconocimiento de ADN se da a través de interacciones del beta-hairpin en el surco mayor generando un *bending* en cada ADN de aproximadamente 25 grados. Las hélices, en cambio, facilitan interacciones hidrofóbicas involucradas por ejemplo en la dimerización.

c) Otros tipos de motivos mixtos alfa/beta

Por ejemplo están presentes motivos mixtos alfa/beta en diversas proteínas con alta especificidad como las endonucleasas. También existen estos motivos en núcleos de otras enzimas como la ADN polimerasa ADN ligasa y otras enzimas de reparación del ADN.

1.3.4: Proteínas multidomino

Existen varias proteínas de unión a ADN cuya interacción viene dada por más de un dominio en la interfase con el ADN, brindando así la posibilidad de reconocer diferentes regiones con variada afinidad y de manera complementaria. Un ejemplo clásico de este tipo de interacciones son las proteínas del dominio POU. En esta familia las proteínas están compuestas por un homeodominio conectado a través de un *linker* flexible a otro de unión específica a POU.

1.4: Consideraciones generales para mecanismos de reconocimiento de proteínas de unión a ADN

A la hora de caracterizar la facilidad con que se pueden formar los complejos es oportuno tener en cuenta dos conceptos importantes: afinidad y especificidad. Por ejemplo, en el caso de complejos ADN-proteína la afinidad viene dada por todos los contactos existentes en la interfase, mientras que la especificidad de la interacción está determinada por aquellos contactos que sean secuencia-específicos. La especificidad en los contactos se da con interacciones a nivel del surco mayor, ya que es la región en la que las bases están accesibles para permitir interacciones de tipo enlaces de hidrógeno o hidrofóbicas. A modo de generalización, puede resumirse entonces que las interacciones específicas a nivel del surco mayor vienen dadas por enlaces de hidrógeno para reconocer específicamente a las purinas (G,A), mientras que interacciones de tipo hidrofóbicas a nivel surco mayor sirven para reconocer de manera específica a las pirimidinas (C,T)

Sin embargo, puede existir especificidad en contactos con ADN que no sean a nivel del surco mayor siempre y cuando los contactos con el *backbone* del ADN se den como causa de una modificación estructural de la molécula debida a la secuencia que la desvíe de la forma B en equilibrio. Existen proteínas, como por ejemplo el represor *Arg*, que reconocen estos cambios estructurales por lo que se puede decir que reconocen estos motivos estructurales mediante un contacto estrecho de una hoja beta sobre el surco menor de manera específica pero indirecta.

Por otro lado, existen muchas proteínas que pueden últimamente estructurarse una vez en presencia del ADN, así como modificar la estructura del mismo al unirse a él. Tal es el caso de algunos leucine zippers que únicamente pueden dimerizar en presencia de ADN, puesto que su dominio básico solamente está estructurado en presencia de ADN.

Una manera de caracterizar las formas de reconocimiento del ADN blanco por parte de las proteínas viene dada por la manera de leer o reconocer al blanco. Se dice que existe un mecanismo de lectura directa, cuando los pares de bases son reconocidos por la proteína

mediante enlaces de hidrógeno así como a través de interacciones hidrofóbicas. En ambos casos pueden darse este tipo de reconocimiento a nivel del surco mayor o menor.

De manera análoga, existen mecanismos de lectura indirecta. En este caso la lectura viene dada por el reconocimiento local y global de modificaciones estructurales, que también puede ser considerado como reconocimiento específico. Por ejemplo, proteínas que reconozcan regiones con *narrowing* en el surco menor, o en general tengan la habilidad de reconocer la forma del mismo, poseen el mecanismo de lectura indirecta. Otro ejemplo dentro de este tipo de mecanismos es la lectura local de forma, como puede ser el reconocimiento de *kinks*. Como ya fue mencionado anteriormente corresponden a regiones donde la orientación del ADN se desvía abruptamente de la linealidad del eje de la hélice. Un ejemplo para este tipo de sistemas es el operón *lac* o la proteína de unión a caja TATA (TBP).
Por otro lado el reconocimiento local también puede producirse al reconocer estructuras particulares de ADN como A-ADN, ADN con *bending* y Z-ADN.

Finalmente, existen también estructuras de complejos ADN-proteína de mayor orden. En este caso se alcanza la formación de un complejo de mayor estabilidad y tamaño como consecuencia de interacciones entre complejos previamente formados. Un ejemplo típico con estas características es el nucleosoma. Aquí las interacciones se dan porque se optimiza la formación de complejos de mayor tamaño una vez que existe el contacto inicial entre ADN y proteína. Es decir, la cromatina se puede formar a causa de que los nucleosomas puedan compactarse más a través de las interacciones de la histona H1 que permite el empaquetamiento de varios nucleosomas y así formar la fibra de 30 nm.

II FUNDAMENTACIÓN METODOLÓGICA

2.1: Simulaciones de dinámica molecular

2.1.1 ¿Por qué usar simulaciones?

La primera pregunta lógica a formularse es ¿Por qué se escogió un método de modelado computacional ante el interés de estudiar un problema biológico? ¿Es un enfoque científico válido? Si bien existe una gran variedad de experimentos que comúnmente se realizan en el laboratorio, se deben dar las condiciones apropiadas para poder estudiar interacciones intermoleculares. Además existen limitaciones en cuanto a la escala y al detalle en que se obtiene la información, así como también de los costos en tiempo y dinero asociados.

El uso de simulaciones de dinámica molecular, introducido en los años 60, es cada vez más utilizado. Hoy en día, su uso conjunto con datos experimentales permite dar mayores detalles y profundizar la comprensión de distintos fenómenos bajo estudio, particularmente en referencia a problemas que se puedan resolver a nivel atómico/molecular. En este campo se han obtenido resultados interesantes de sistemas biológicos cuyos detalles estructurales fueron resueltos con métodos computacionales como las simulaciones de dinámica molecular. Un ejemplo ilustrativo de la utilidad de estos métodos es la descripción atómica del proceso de separación de las dos hebras del ADN mediante simulaciones el cual es difícil de estudiar experimentalmente (Perez and Orozco 2010.) Otros sistemas muy estudiados por métodos computacionales son aquellos que involucran proteínas de membrana con la finalidad de elucidar los mecanismos de acción de dichas proteínas a nivel mecánico en presencia de la membrana, así como también el interés en modelar proteínas que actúen como motores moleculares. Este es un campo de gran interés debido a la dificultad de obtener esta información por métodos experimentales. Existen varios ejemplos de este tipo de sistemas, como el de la ATP sintetasa (también conocida como F_oF_1-ATP sintasa) en el que por ejemplo se ha podido analizar los efectos estructurales del complejo en función de la rotación de la subunidad γ (Karplus and Kuriyan 2005) Otro ejemplo más reciente muestra el poder de cálculo cada vez mayor que existe en esta área, permitiendo haber estudiado la dinámica de la cápside completa del virus de VIH-1 a partir de su estructura resuelta por crio-electromicroscopía(Zhao, Perilla et al. 2013). Este sistema compuesto por aproximadamente 64 millones de átomos permitió estudiar como las proteínas de la cápside al ensamblarse en diferentes motivos estructurales, como pentámeros de hexámeros o hexámeros de hexámeros, afecta la curvatura de la cápside, aportando así nociones del mecanismo de ensamblado viral.

En los siguientes capítulos de la introducción se pretende dar una noción básica de los principales fundamentos teóricos vinculados con la técnica de simulaciones por dinámica molecular.

2.1.2 Algunas generalidades

Existen diferentes métodos teóricos para estudiar estructuras moleculares y todos están fundamentados en algún punto en la ecuación de Schrödinger que describe la vinculación entre la función de onda de un sistema y su energía. La ecuación de Schrödinger independiente del tiempo, puede escribirse en su forma más desarrollada como:

$$E\Psi = -\frac{\hbar^2}{2\mathrm{m}}\nabla^2\Psi + \hat{V}(r)\Psi \qquad (2.1)$$

Donde $\hbar = \frac{h}{2\pi}$ es la constante de Planck reducida, $\hat{V}(r)$ corresponde al término de la energía potencial de las partículas en función de su posición y Ψ a la función de onda. Asimismo la ecuación de Schrödinger independiente del tiempo (2.1) puede representarse de forma más reducida de la siguiente manera:

$$\hat{H}\Psi = E\Psi \qquad (2.2)$$

Donde $\hat{H}$ corresponde al operador Hamiltoniano ($\hat{H} = \frac{\hbar^2}{2\mathrm{m}}\nabla^2 + \hat{V}$), Ψ a la función de onda y E a la energía del sistema (potencial y cinética).

Sin embargo, puntualmente para el caso de simulaciones de dinámica molecular donde el objetivo es describir un sistema molecular en función de su evolución temporal y los sistemas de interés poseen una gran cantidad de átomos (generalmente interesa simular comportamientos de macromoléculas) es que se utiliza un enfoque clásico para modelar las interacciones existentes y consecuentemente poder integrar las ecuaciones de movimiento de Newton. Esta estrategia puede ser implementada como consecuencia de algunas aproximaciones realizadas a detallarse a continuación. La aproximación de Born-Oppenheimer es la primera de ellas. En ella se plantea la posibilidad de disociar los movimientos nucleares de los electrónicos ya que la masa electrónica es mucho menor que la nuclear y las energías asociadas a ambos son considerablemente distintas. Por este motivo, las variaciones en las posiciones nucleares ocurren a una escala temporal mucho mayor comparada con las electrónicas, teniendo entonces los electrones la capacidad de adaptarse rápidamente a estos cambios nucleares, pudiendo considerarse que la adaptación a dichos cambios es casi instantánea. Esta posibilidad de separar la descripción del movimiento electrónico del nuclear es la que permite realizar la segunda aproximación donde se modela el comportamiento de los núcleos de una manera clásica (es decir, cumplen a las leyes clásicas de la física). Lo que nos lleva a introducir la última aproximación utilizada, la base de la obtención de las energías en mecánica clásica: los campos de fuerza, que pueden ser empíricos o con información derivada de la mecánica cuántica, donde se considerarán los efectos electrónicos.

2.1.3 Teoría de las simulaciones de dinámica molecular

La dinámica molecular es una técnica computacional que se basa en integrar las ecuaciones clásicas de movimiento para los átomos del sistema en cada paso temporal. Es decir que se resuelve la ecuación de movimiento proveniente de la física clásica, que en el caso de un átomo tiene la siguiente expresión:

$$F_i = m_i \cdot a_i = m_i \cdot \ddot{r}_i \tag{2.3}$$

Donde F_i corresponde a la fuerza actuante sobre un átomo *i* como consecuencia de la interacción con el resto de los átomos, m_i a la masa de dicho átomo y a_i a la aceleración (o lo que es lo mismo, la segunda derivada de la posición con respecto al tiempo, $\ddot{r}_i$).

La ecuación (2.3) es una ecuación diferencial de segundo orden, que puede reescribirse como dos ecuaciones diferenciales de primer orden, una expresando la derivada del momento y la otra la derivada de las posiciones de los átomos, es decir:

$$\dot{p}_i = \frac{\partial p_i}{\partial t} = F_i \tag{2.4}$$

$$\dot{r}_i = \frac{\partial r_i}{\partial t} = \frac{p_i}{m} \tag{2.5}$$

donde p_i es el momento de un átomo i ($p_i = m_i \cdot v_i$), r_i su posición y F_i la fuerza total que actúa sobre dicho átomo. A su vez, F_i puede expresarse como la sumatoria de fuerzas de los pares de átomos no unidos más las fuerzas existentes entre átomos unidos (que pueden involucrar 2, 3 o 4 átomos) más las fuerzas externas o de restricción que puedan existir. Estas fuerzas a su vez, pueden resolverse desde el gradiente de la energía potencial causada por desplazamientos de los átomos:

$$F_i(r_1,..r_N) = -\nabla_{r_i} E_{pot}(r_i,..,r_N) \tag{2.6}$$

Finalmente, podemos combinar (2.3) con (2.6) y obtener la siguiente expresión de la ecuación de movimiento para un sistema de N átomos:

$$-\nabla E_{pot}(r_1,..r_N) = m_i \frac{d^2 r_i}{dt^2} \tag{2.7}$$

Es precisamente en este punto, donde los cálculos están basados en expresiones de la energía potencial $E_{pot}(r)$, que es oportuno introducir el concepto de campo de fuerza.

2.1.4 Campos de fuerza (*Force fields*)

En química computacional, un campo de fuerza es un conjunto de parámetros y funciones usados para hallar la energía potencial de un sistema. Estos parámetros pueden obtenerse desde cálculos de mecánica cuántica o en base a datos experimentales. Generalmente se obtienen los datos para moléculas pequeñas o grupos funcionales que luego se transfieren a moléculas mayores.

Si bien existen variaciones en la forma de modelar la energía potencial de un sistema atómico, los campos de fuerza más ampliamente utilizados actualmente han convergido en el uso de un conjunto de términos que representan los modos de interacción interatómicas, como se muestra en la tabla 2.1.

Tabla 2.1: Lista de interacciones básicas que se expresan en los campos de fuerza

Tipo de interacción	Expresión matemática para la energía potencial
Longitud del enlace ("bond stretching")	$\sum_{enlace} \frac{1}{2} k_r (r - r_0)^2$
Ángulo ("angle bending")	$\sum_{ángulo} \frac{1}{2} k_\theta (\theta - \theta_0)^2$
Ángulo del diedro (torsión)	$\sum_{torsión} \frac{U_n}{2} (1 + \cos(n\varphi - \delta))$
Interacción electrostática (Coulomb)	$\sum_{Coulomb} k_{el} \left(\frac{q_i q_j}{r_{ij}}\right)$; $k_{el} = \frac{1}{4\pi\varepsilon}$
Interacción de Van der Waals: potencial de Lennard-Jones	$\sum_{L-J} \left(\frac{A_{ij}}{r_{ij^{12}}} - \frac{B_{ij}}{r_{ij^6}}\right)$; $4\varepsilon\sigma^{12} = A$; $4\varepsilon\sigma^6 = B$
Expresión de la energía potencial total:	$U = \sum_{enlace} + \sum_{ángulo} + \sum_{torsion} + \sum_{L-J} + \sum_{Coulomb}$

Estos términos son las expresiones mínimas que posee cualquier campo de fuerza y pueden diferenciarse en dos grandes grupos: las interacciones enlazantes y las no enlazantes. Dentro del primer grupo se encuentran todas aquellas interacciones que se dan como consecuencia de la unión covalente entre átomos. En dicho grupo existen expresiones para las energías aportadas como consecuencia de la vibración de los enlaces (interacciones que involucran 2 átomos), variación del ángulo entre 3 átomos enlazados entre sí y la variación de ángulos torsionales (4 átomos consecutivos involucrados).

a.1) Potenciales de enlace: “bond stretching”

El término corresponde al potencial existente como consecuencia de la variación de la longitud del enlace (estiramiento o acortamiento) respecto a la de referencia, siendo la longitud de referencia aquella cuando todos los restantes términos del campo de fuerza son nulos, que no corresponde necesariamente a la longitud de equilibrio del enlace. En mecánica molecular se suele utilizar la expresión escrita en la tabla 2.1, donde se expresa en forma de un simple potencial armónico de Hooke siendo k_r la constante elástica del resorte, r la distancia entre los átomos que participan del enlace y r_0 la distancia de equilibrio entre ambos átomos.

a.2) Potenciales de enlace: “angle bending”

Este tipo de potencial usualmente se formula mediante una expresión armónica al igual que en el caso anterior, con la salvedad de que el potencial se aplica entre dos átomos conectados entre sí mediante un átomo central ligado covalentemente a ambos. En la tabla 2.1 k_θ corresponde a la constante elástica, θ al ángulo definido por los tres átomos y θ_0 al ángulo de equilibrio entre los 3 átomos.

a.3) Potenciales de enlace: Diedros torsionales

Describe la posición relativa de dos átomos separados por tres enlaces covalentes, mediante el desarrollo en serie de coseno como se ve en la tabla 2.1, donde U_n es la altura de la barrera, φ es el ángulo del diedro definido por los cuatro átomos, n la multiplicidad y δ la fase.

b.1) Potencial electrostático: Potencial de Coulomb

Otro término importante a contemplar son las contribuciones electrostáticas a la energía del sistema a modelar. Este tipo de interacción no enlazante se suele modelar generalmente como potenciales de Coulomb entre pares de átomos. De acuerdo a la expresión de la tabla 2.1 k_{el} es la constante de Coulomb, q_i y q_j las cargas de los átomos i y j respectivamente y r_{ij} la distancia entre los átomos i y j. Un dato importante a tener en cuenta es el radio de acción de dichas interacciones, que será una de las razones principales en el costo computacional de los cálculos. Es aquí entonces que existen distintas técnicas para considerar la distancia máxima en que dichas interacciones (*cut-off*) así como métodos para evitar problemas de borde (típicamente causados por discontinuidades en la derivada en el punto de borde) (Schlick 2010).

b.2) Potencial de Lennard-Jones

Las interacciones de Van der Waals son el otro tipo de interacciones no enlazantes que dan estabilidad a la estructura de las moléculas y también se calculan considerando pares de

átomos. Estas interacciones incluyen interacciones de tipo dipolo-dipolo, dipolo-dipolo inducido e interacciones de dispersión de London (dipolo inducido instantáneo-dipolo inducido). Una forma de incluirlas dentro de los cálculos de mecánica molecular es como potenciales de tipo Lennard-Jones, donde es posible considerar las dos características importantes de este tipo de interacciones no enlazantes. Estas son atractivas a largo alcance y repulsivas a corto alcance. Por ejemplo, expresando el término de Lennard-Jones de acuerdo a la tabla 2.1; donde ε es la profundidad del pozo y σ el diámetro de colisión (separación interatómica tal que la energía vale cero); el término r^{-12} modela la a repulsión de Pauli, mientras que el r^{-6} modela las atracciones debida a fuerzas de Van der Waals. Para calcular las interacciones de van der Waals entre los pares de átomos suelen utilizarse reglas de combinación. Una de las más utilizadas, y la que se usará en el presente trabajo, se denomina Lorentz-Berthelot. Ésta calcula el parámetro sigma de la interacción como la media aritmética y el épsilon como la media geométrica de los valores de sigma y épsilon del par de átomos respectivamente. Es decir para una interacción de vdW entre un átomo i y j la expresión para σ y ε se obtendrá de la siguiente manera:

$$\pm_{ij} = \frac{\pm_i + \pm_j}{2} \qquad £_{ij} = \sqrt{£_i £_j}$$

Dependiendo de las propiedades e implementación de los campos de fuerza que se consideren, pueden existir variaciones en las formas de expresar los potenciales así como también en los términos que se puedan agregar. Un ejemplo de esto son aquellos términos extra utilizados históricamente para modelar con mayor detalle los enlaces de hidrógeno (pudiendo agregar un término específico para ello), las interacciones metal-ligando, efectos de solvatación, etc.

2.1.5 Algoritmos utilizados en dinámica molecular

Existen varios métodos para integrar las ecuaciones de movimiento newtonianas. En todos ellos se debe considerar un paso temporal (*timestep*) suficientemente pequeño para asegurar la conservación de la energía. Por este motivo los movimientos de mayor frecuencia serán los que limiten el valor de *timestep* a utilizar. En el caso de las moléculas biológicas éstos corresponden a los movimientos de vibraciones de los enlaces que contengan átomos de hidrógeno, debido a la pequeña masa de dichos átomos. A continuación se presentan brevemente tres de los métodos más comunes para la propagación del movimiento en dinámica molecular y la relación que tienen entre sí (Leach 2001).

Verlet

Este algoritmo de integración es el más simple de entender y se basa en obtener las posiciones a tiempo t a tiempo (t-δt) y la aceleración a tiempo t. La aceleración se obtiene

directamente a través de la evaluación de la fuerza de la partícula i considerada del siguiente modo:

$$a_i(t) = \ddot{r}(t) = \frac{d^2 r_i}{dt^2} = \frac{f_i(t)}{m_i} \tag{2.8}$$

Por otro lado, las posiciones se obtienen a través del desarrollo de Taylor de la función que representa la posición en función del tiempo t-δt y t+δt. truncadas en el segundo término, correspondiendo el resto al error del método:

$$\left.\begin{aligned} r(t+¢t) &= r(t) + ¢t v(t) + \frac{1}{2}¢t^2 a(t) + O¢t^4 \\ r(t-¢t) &= r(t) - ¢t v(t) + \frac{1}{2}¢t^2 a(t) + O¢t^4 \end{aligned}\right\}$$ sumando ambas se obtiene la expresión del algoritmo de Verlet:

$$r(t+¢t) = 2r(t) - r(t-¢t) + ¢t^2 a(t) \tag{2.9}$$

Una vez que se obtienen las posiciones se calculan las velocidades en función de ellas, ya que no aparecen explícitamente en la ecuación anterior. Una forma para calcularla consiste simplemente en obtener la velocidad como el cociente entre la diferencia entre las posiciones a tiempo t+δt y t-δt y 2δt es decir:

$$v(t) = \frac{r(t+¢t) - r(t-¢t)}{2¢t} \tag{2.10}$$

Velocity-verlet

Una variante del algoritmo de Verlet es el denominado “Velocity-Verlet”, donde las aceleraciones, posiciones y velocidades se calculan todas para el mismo tiempo de acuerdo a las siguientes expresiones:

$$r(t+¢t) = r(t) + ¢t v(t) + \frac{¢t^2}{2} a(t) \tag{2.11}$$

$$v(t+¢t) = v(t) + \frac{¢t}{2}(a(t) + a(t+¢t)) \tag{2.12}$$

Para continuar con la dinámica es preciso calcular las velocidades a tiempo t+δt/2 que se obtienen de la siguiente forma:

$$v(t+\frac{¢t}{2}) = v(t) + \frac{¢t}{2} a(t) \tag{2.13}$$

Luego las nuevas fuerzas se calculan desde las posiciones actuales, dando como resultado a(t+δt) y finalmente, en el último paso se determinan las velocidades para el tiempo t+δt de acuerdo a:

$$v(t+¢t) = v(t+\frac{¢t}{2}) + \frac{¢t}{2} a(t+¢t) \tag{2.14}$$

Leap-frog

Este algoritmo lleva el nombre de “”leap-frog” (en español: “salto de rana”) debido a que se obtienen las velocidades y las posiciones en tiempos diferentes de manera solapada. La velocidad se calcula cada medio paso de tiempo (δt/2) mientras que las posiciones lo hacen a cada paso de tiempo (δt).

Las posiciones y las velocidades se obtienen de los valores a tiempo t y δt/2 respectivamente. Las mismas se actualizan considerando la aceleración a del mismo modo que sucede en Verlet (ver 2.10).

$$v(t+\frac{¢t}{2}) = v(t-\frac{¢t}{2}) + a(t)¢t \tag{2.15}$$

$$r(t+¢t) = r(t) + v(t+\frac{¢t}{2})¢t \tag{2.16}$$

Cuando se implementa el algoritmo de leap-frog las velocidades $v(t+\frac{t}{2})$ se calculan inicialmente de las velocidades a tiempo $v(t-\frac{¢t}{2})$ y de las aceleraciones a tiempo t.

2.1.6 Control térmico y de presión durante las simulaciones

Durante una simulación de dinámica molecular se puede trabajar con diferentes tipos de ensambles termodinámicos, como por ejemplo, el canónico($\mathcal{N},\mathcal{V},\mathcal{T}$), microcanónico ($\mathcal{N},\mathcal{V},\mathcal{E}$) o como generalmente sucede en los laboratorios, trabajar con el ensamble isobárico-isotérmico ($\mathcal{N},\mathcal{T},\mathcal{P}$). Para asegurarse que se trabaja en el tipo de ensamble deseado es que es necesario utilizar métodos de control de aquellas magnitudes constantes a lo largo de toda la simulación.

Existen diferentes métodos para mantener la presión constante a lo largo de la simulación, uno de ellos es el de Berendsen (escogido para la termalización inicial en este trabajo), donde se acopla débilmente el sistema a un baño externo usando el principio de mínima perturbación local. Esto se logra agregando un término extra a la ecuación de movimiento que efectúe los cambios necesarios para mantener constante la presión.

$$\left(\frac{\partial p}{\partial t}\right)_{baño} = \frac{p_0 - p}{\tau_p} \tag{2.17}$$

Siendo τ_p la constante del tiempo para el acoplamiento

Análogamente, para asegurar que la temperatura promedio del sistema sea correcta también se agrega un término más a la ecuación de movimiento. Una forma de hacer esto es a través del termóstato de Berendsen donde se eliminan las fluctuaciones en la energía cinética de modo de asegurar que la temperatura se mantenga estable (Berendsen, Postma et al. 1984). Las velocidades se escalan entonces a cada paso de modo que la tasa de cambio de la

temperatura sea proporcional a la diferencia entre la temperatura del baño acoplado y la del sistema de la siguiente forma:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{baño} - T(t)) \qquad (2.18)$$

Donde τ representa la constante de acoplamiento entre el baño y el sistema, a mayor τ, mayor será el acoplamiento.

2.1.7 Algoritmos para minimización energética

Una vez obtenida una expresión que permita calcular la energía de un sistema molecular dado, es posible recorrer la superficie de energía potencial que esta expresión describe. Los mínimos de esta superficie multidimensional representan las conformaciones energéticamente más estables del sistema estudiado.

Para realizar esto existen diversos métodos, donde el más apropiado dependerá del sistema en el que se trabaje, así como del enfoque (cuántico o mecánico clásico), considerando siempre la relación del gasto computacional respecto a la calidad del resultado. Dos de los algoritmos más utilizados son los llamados: "steepest descent" y "conjugate gradient". A modo informativo es conveniente explicar un poco en qué consisten cada uno de estos algoritmos.

Descenso por gradiente (*steepest descent)*

Este método derivativo de primer orden se basa en un algoritmo que explora la superficie de energía potencial moviéndose de un paso a otro en la dirección paralela a la fuerza neta. La fuerza neta se calcula mediante el gradiente de la energía potencial ($F=-\nabla E_{pot}$) de la estructura inicial de acuerdo con el campo de fuerza elegido. Como se trata de un método iterativo cada nueva posición generada en un paso determinado se obtiene a partir de la posición resultante del paso anterior. Esquemáticamente el algoritmo seguido es el siguiente:

- Dado un punto inicial en la superficie de energía potencial se calcula el gradiente de ella en ese punto, definiendo así la dirección en la que se moverá sobre la superficie:

$s_k = -\frac{g(x_k)}{|g(x_k)|}$ donde s_k es el vector unitario de dirección de búsqueda, x_k las coordenadas en el paso k y $g(x_k)$ el gradiente de la energía potencial.

- Con la dirección definida (s_k) se calcula la distancia en que se moverá:

$x_{k+1} = x_k + \lambda_k s_k$ donde, λ_k es la magnitud del salto en la dirección de búsqueda.

λ_k puede ser un valor arbitrario fijo asignado u obtenerse mediante una búsqueda lineal. En este último caso el objetivo es identificar el mínimo a lo largo de una dirección específica (de allí el nombre en inglés *"linear search"*).

Estos pasos se repiten hasta que se alcance alguna de las condiciones finales dadas por el usuario. Puede ser hasta alcanzar un número máximo de pasos o al alcanzar un valor de gradiente menor a un valor especificado.

Gradiente conjugado (conjugate gradient)

El método de gradiente conjugado también es derivativo de primer orden, pero el algoritmo es diferente al del caso anterior. Esto se debe a que se utiliza otro criterio para escoger la dirección en la que se moverá sobre la superficie de energía potencial. Como lo indica el nombre, la dirección se obtiene de conjugar las dos direcciones de los gradientes de los pasos consecutivos*:*

$$v_k = -g(x_k) + \gamma_k v_{k-1}$$

donde v_k es el vector dirección de búsqueda en el paso k, x_k las coordenadas en el paso k, $g(x_k)$ el gradiente de la energía potencial en el paso k y γ_k es un escalar definido en su versión original (algoritmo de Fletcher-Reeves) como:

$$\gamma_k = \frac{g(x_k) \cdot g(x_k)}{g(x_{k-1}) \cdot g(x_{k-1})}$$

Si bien conjugate gradient es computacionalmente más costoso que steepest descent, este método converge más rápidamente que el anterior y comúnmente se lo utiliza a continuación de steepest descent.

2.2: Modelos *Coarse-grained*

2.2.1 Generalidades de modelos de grano grueso (*Coarse-grained*)

Los modelos simplificados de grano grueso, llevan adquiriendo mayor relevancia en los últimos años en la comunidad de biofísica computacional debido a la alta demanda de recursos que la mayoría de los sistemas de interés biológico necesitan para ser estudiados en escalas temporales y espaciales de interés. En este tipo de enfoque la idea general consiste en reducir los grados de libertad del sistema, reduciendo así el tamaño del sistema a estudiar (Noid 2013). Por este motivo es importante tener en cuenta las características específicas del sistema de manera de poder representar de manera correcta los procesos de interés a pesar de la simplificación en el mismo. Dentro de la clasificación de sistemas de grano grueso (en inglés:

coarse-grained) se pueden dividir varias aproximaciones. Existen enfoques de tipo red elástica (elastic network) donde una proteína se modela como una serie de esferas conectadas a través de elásticos (con una constante armónica asociada) permitiendo que ésta pueda variar su conformación entorno al mínimo energético de dicha red. Este es un enfoque que se utiliza por ejemplo para reconocer dinámicas de motivos estructurales que no se desvíen demasiado de la estructura de referencia. (Periole, Cavalli et al. 2009)

Del mismo modo otra forma de simplificar el sistema a simular mediante dinámica molecular, es el denominado *coarse-grained* basado en partículas. En este caso se simplifica la representación de los componentes del sistema, por ejemplo de una proteína, por centroides (o pseudoátomos) con propiedades adecuadamente escogidas para que representen un grupo de átomos que estaría explícitamente considerado en una dinámica atomística común. Un ejemplo minimalista de este tipo de simplificación es por ejemplo un campo de fuerzas de tipo "united atoms" como gromos. En este campo de fuerza se representan algunos hidrógenos implícitamente con el carbono al que están unido covalentemente (como lo es el caso de los carbonos CH2). Sin embargo, existen verdaderos modelos *coarse-grained* donde se pueden representar a un grupo de átomos con una esfera como lo son OPEP, Martini, PaLaCe, etc(Marrink, Risselada et al. 2007; Monticelli, Kandasamy et al. 2008; Baaden and Marrink 2013; Sterpone, Melchionna et al. 2014) PaLaCe por ejemplo, utiliza un modelo CG de dos niveles donde la región del enlace peptídico se representa por interacciones atomísticas mientras que las cadenas laterales de cada aminoácido son modeladas por pseudoátomos mediante interacciones de vdW y eventualmente algún átomo atomístico para reproducir interacciones como enlaces de hidrógeno (Pasi, Lavery et al. 2013).

2.2.2 Generalidades del campo de fuerza SIRAH

Este campo de fuerza ha sido desarrollado enteramente en el grupo de Simulaciones Biomoleculares del Institut Pasteur de Montevideo y debe su nombre al acrónimo en inglés *Southamerican Initiative for a Rapid and Accurate Hamiltonian* (Darré, Machado et al. 2015) Inicialmente se desarrollaron parámetros para ADN doble hebra, seguido del solvente explícito *co*arse-grained y finalmente proteína. En todos estos casos se ha seguido una estrategia de *coarse graining* basado en partículas, donde se trabaja con topologías y parámetros de interacción para los átomos *coarse-grained* (CG) representados pero manteniendo las expresiones generales de un Hamiltoniano atomístico (tabla 2.1). Esto implica que haya expresiones para la energía potencial de enlaces, ángulos, diedros, diedros impropios, interacciones de tipo Lennard Jones y electrostática tal como sucede en la mayoría de los campos atomísticos (ver sección 2.1.4). La única diferencia es que estos términos expresan ahora las interacciones presentes en las partículas del campo de fuerza simplificado en vez de entre átomos reales. Es importante destacar que los átomos CG siempre se mapean desde posiciones atomísticas, es decir, el centro de un átomo CG siempre está ubicado sobre una posición correspondiente a un átomo real (también denominados FG, del inglés *Fine Grain*) existente.

Tabla 2.2: Mapeo y parámetros de SIRAH. En todos los casos se presenta la estructura atomística únicamente a través de átomos pesados y aquellos hidrógenos desde los que se mapea a SIRAH. En el caso de aminoácidos los átomos numerados del 1-3 están presentes en todos ya que corresponden al *backbone* y se muestran explícitamente para glicina y alanina solamente.

FG	CG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)	FG	CG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)
G		1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,40 0,40 0,40	0,55 0,55 0,55	A		1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,41 0,40 0,40	2,00 0,55 0,55
S		4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	I		4: BCG	0	0,41	3,20
T		4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	V		4: BCB	0	0,41	3,20
N		4: BCG 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	L		4: BCG	0	0,41	3,20
Q		4: BCD 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	C		4: BSG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01
Y		4: BCG 5: BCE1 6: BCE2	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	M		4: BSD	0	0,45	3,20
He		4: BCG 5: BNE 6: BND	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	P		4: BCG	0	0,43	0,60
K		4: BCG 5: BCE	0,40 0,60	0,40 0,55	0,55 0,55	F		4: BCG 5: BCE1 6: BCE2	0 0 0	0,35 0,35 0,35	1,70 1,70 1,70
R		4: BCG 5: BCZ 6: BNN1 7: BNN2	0 0,30 0,35 0,35	0,40 0,40 0,45 0,45	0,55 0,35 0,55 0,55	W		4: BCG 5: BNE 6: BPE 7: BCZ 8: BCE	0 -0,10 0,10 0 0	0,35 0,35 0,35 0,35 0,35	1,70 0,10 0,01 1,70 1,70
D		4: BCG 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55	E		4: BCD 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55
A		1: PX 2: C5X 3: C1A 4: N6A 5: N1A 6: C2A	-1 0 0 0,2 -0,2 0	0,46 0,43 0,34 0,32 0,32 0,27	0,84 0,46 0,46 1,00 1,00 0,46	G		1: PX 2: C5X 3: C1G 4: O6G 5: N1G 6: N2G	-1 0 0 -0,4 0,2 0,2	0,46 0,43 0,34 0,30 0,32 0,32	0,84 0,46 0,46 1,30 1,09 1,09
T		1: PX 2: C5X 3: C1T 4: O4T 5: N3T 6: O2T	-1 0 0 -0,2 0,4 -0,2	0,46 0,43 0,34 0,30 0,32 0,30	0,84 0,46 0,46 1,00 0,80 1,00	C		1: PX 2: C5X 3: C1C 4: N4C 5: N3C 6: O2C	-1 0 0 0,4 -0,2 -0,2	0,46 0,43 0,34 0,32 0,32 0,30	0,84 0,46 0,46 1,09 1,09 1,30
K^+ and 6 water molecules		1: KW	1,00	0,645	0,55	WT4 11 water molecules		1: WN1 2: WN2 3: WP1 4: WP2	-0,41 -0,41 0,41 0,41	0,42 0,42 0,42 0,42	0,55 0,55 0,55 0,55
Na^+ and 6 water molecules		1: NaW	1,00	0,58	0,55	Cl^- and 6 water molecules		1: ClW	-1,00	0,68	0,55

La elección de los átomos CG fue basada en puntos de interacción importantes y característicos de cada aminoácido. En la tabla 2.2 se puede observar por ejemplo, que aquellos aminoácidos polares donde la existencia de un hidroxilo es importante para las interacciones de su cadena lateral mediante enlaces de hidrógenos fue modelado con la presencia explícita de átomos CG con cargas parciales en la posición del O y del H de dicho grupo funcional. Si bien la granularidad es mayor en este modelo, es posible mantener un efecto de dipolo en dicho grupo funcional lo que le da características específicas a dicha cadena lateral. También es visible que los aminoácidos básicos y ácidos se encuentran representados con una carga neta en la cadena lateral acorde a la protonación a pH 7 de los mismos. En el caso de la histidina, existen parámetros tanto para la protonación en la posición δ o ε del imidazol.

En el caso del ADN, los *beads* centrales de las bases poseen parámetros de van der Waals análogos a los existentes en el campo de fuerza atomístico AMBER, permitiendo así que haya un correcto empaquetamiento del ADN al estar el stacking estéricamente permitido. Esto es importante ya que cada nucleótido está representado por 6 átomos, 3 correspondientes a la base nitrogenada, 2 correspondientes a los átomos 1´ y 5' de la desoxirribosa y el último representando al fosfato (ver tabla 2.2).

A su vez, el *backbone* está representado con átomos que poseen parámetros de van der Waals como los existentes en el campo de fuerza atomístico AMBER y las cargas en el caso de las proteínas representan un potencial electrostático análogo al de una alfa hélice con parámetros de AMBER. Esto permite que exista una correcta representación de estructura secundaria sin necesidad de restricciones adicionales para mantenerla.

Por otra parte, el solvente coarse-grained explícito, WT4, también tiene sus orígenes en el mapeo de posiciones reales de átomos. En este caso, dado que el agua es un líquido estructurado, se mapean posiciones correspondientes a oxígenos en cada vértice de un cluster de agua. Este modelo permite representar la característica estructurada del agua siguiendo las características atomísticas pero aumentando la granularidad del sistema. La distancia entre cada pseudoátomo de WT4 hacia el centro del tetraedro corresponde a 2.8 Å, que es la distancia inter-oxígenos en la primera capa de solvatación del agua. Por este motivo, es que se considera que WT4 es capaz de "solvatar" explícitamente en la segunda capa de solvatación únicamente en caso de no haber modificaciones específicas de la interacción entre WT4 y el pseudoátomo en cuestión. Esto también va en línea con el modelo de iones monovalentes existentes, donde se modela al mismo considerando su primera esfera de solvatación de manera implícita (Darré, Machado et al. 2010).

Es importante remarcar que la expresión de la energía potencial total de un sistema *coarse-grained* en SIRAH es la misma que la presentada en la tabla 2.1 en la sección de

fundamentación metodológica. Sin embargo, en algunos casos resulta importante tener más detalle en la descripción de la interacción entre algunos tipos de átomos. Este es el caso de los átomos del *backbone* de la proteína por ejemplo, donde para permitir una correcta descripción de la estructura secundaria de una alfa hélice, es necesario que tengan tamaños atomísticos. Sin embargo, en interacciones de largo alcance ese nivel de descripción no es necesario. Para permitir que un mismo átomo interaccione con diferentes características dependiendo de con que partícula interaccione, existe la posibilidad de establecer explícitamente valores fijos para sigma y épsilon de la interacción entre los pares de átomos. Es decir, que para ciertos pares de átomos es posible definir interacciones fuera de las reglas de combinación de Lorentz-Berthelot. En la figura 1 se muestra una matriz donde se representan todas las posibles combinaciones de interacciones de van der Waals entre los tipos de átomos presentes en SIRAH. Los puntos en blanco son calculados únicamente por reglas de combinación de tipo Lorentz-Berthelot donde el parámetro sigma de la interacción se obtiene como la media aritmética y el épsilon como la media geométrica de los valores de sigma y épsilon del par de átomos respectivamente.

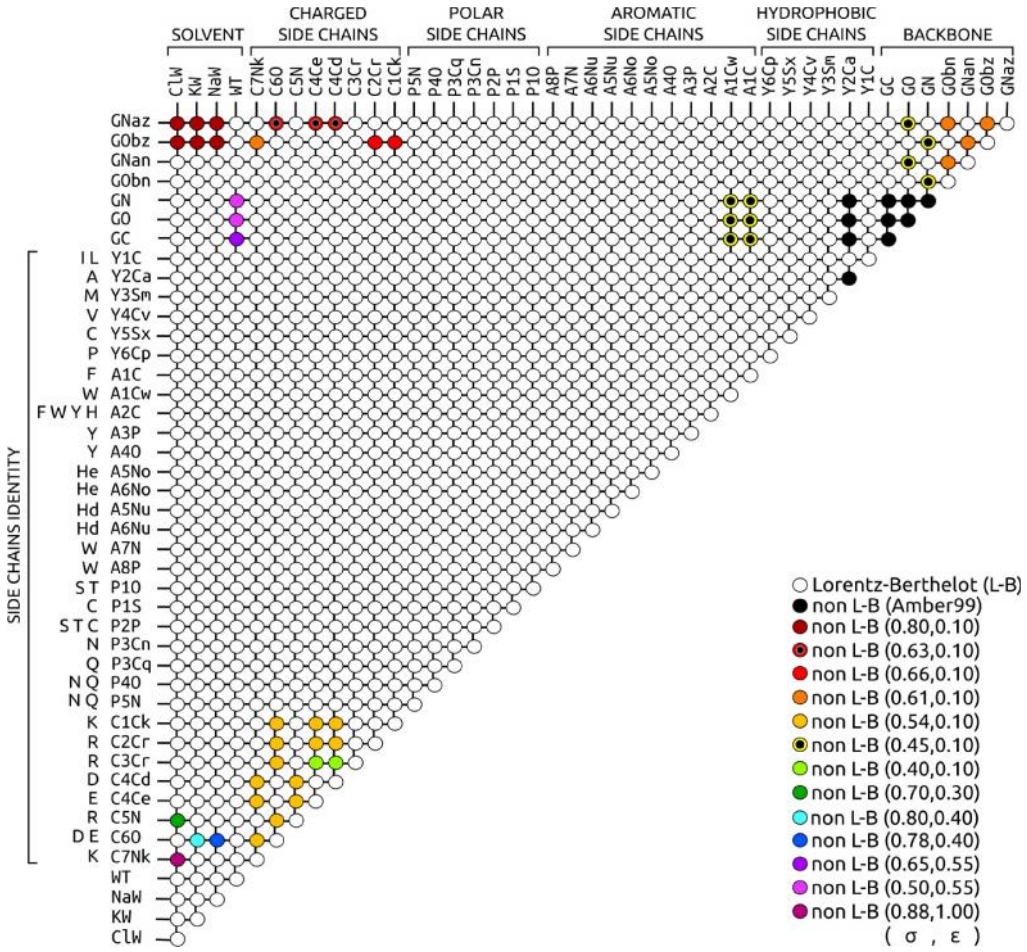


Figura 1: Matriz de interacciones van der Waals entre todos los tipos de átomos presentes en el campo de fuerza. Los círculos blancos son calculados por las reglas de Lorentz-Berthelot, mientras que los coloreados los valores de sigma y épsilon de la interacción fueron seteados específicamente a los pares de valores de acuerdo a la leyenda presente (non L-B) (tomado de Darré, Machado et al. 2015)

Además de las interacciones del *backbone* de la proteína consigo misma existen algunas otras pocas interacciones corregidas, como la de los iones con algunos átomos cargados de lisina, arginina, glutamato o aspartato para evitar interacciones electrostáticas artificiales debido a la escasa competencia del solvente CG por interacciones intramoleculares. Este efecto es intrínseco a las características supramoleculares y debe considerarse una limitación del modelo. En términos generales, esto implica que los parámetros de van der Waals pueden ser fácilmente utilizados para modular las interacciones. Esta estrategia será utilizada para parametrizar las interacciones entre ADN y proteína y corregir interacciones espurias que pudieran existir.

La ganancia computacional promedio del uso de este campo de fuerza está entorno a un factor de 100 en comparación con simulaciones atomísticas de los mismos sistemas realizados en GROMACS4.5.5. Esta ganancia viene dada tanto por la reducción del número de partículas del sistema, como por el aumento del paso de tiempo en SIRAH que es de 20 fs.

III DETALLES COMPUTACIONALES

3.1 Preparación de sistemas

Las coordenadas para cada uno de los sistemas fueron obtenidas de los archivos existentes en la base de datos de sistemas resueltos experimentalmente, PDB. En caso de ser estructuras resueltas por cristalografía, dichas estructuras fueron protonadas a ph 7 con el servidor pdb2pqr (http://nbcr-222.ucsd.edu/pdb2pqr_1.8/)(Dolinsky, Nielsen et al. 2004; Bussi, Donadio et al. 2007) utilizando el esquema de nombres de AMBER para luego ser transformadas con un script *ad hoc* a *coarse-grain* (CG).

3.2 Protocolo de Simulación

Todos los sistemas fueron simulados con Gromacs (v4.5.5)(Pronk, Páll et al. 2013) con los parámetros correspondientes al campo de fuerza *coarse-grained* SIRAH y las modificaciones para ADN-proteína que correspondieren al caso. Estas simulaciones fueron hechas con solvente explicito (WT4) e iones CG que representan una concentración aproximada de 150 mM KCl o NaCl. Se llevaron a cabo 1000 pasos de minimización con algoritmos de steepest descent. Luego se equilibró con restricciones de posición calentando hasta 300 K a presión constante (1 bar) isotrópica, utilizando el termóstato V-rescale(Bussi, Donadio et al. 2007) y baróstato de Parinello-Rahman(Parrinello and Rahman 1981). En los sistemas con proteínas completas se realizó un paso extra de equilibración siguiendo el protocolo anteriormente mencionado durante 100ns con restricciones de posición en el ADN únicamente, permitiendo la relajación de la proteína en la interafse con el ADN. Las corridas de producción se realizaron por 1 μs trabajando con condiciones periódicas de contorno, así como con tratamiento de tipo “Particle Mesh Ewald” para las interacciones electrostáticas. El valor de cutoff correspondió a 12 Å para las interacciones de largo alcance (vdW y electrostáticas). Se utilizó un paso de integración de 20 fs y la información fue guardada cada 100 ps. Las dinámicas fueron visualizadas con VMD (Humphrey, Dalke et al. 1996; Shui, McFail-Isom et al. 1998).

3.3 Análisis

Para calcular el área de interfase entre ADN y proteína se obtuvieron los valores de superficie accesible al solvente mediante la herramienta g_sas de Gromacs, utilizando un radio de solvente de 0.21 nm correspondiente a WT4 y los valores de radios de van der Waals correspondientes al campo de fuerza utilizado (SIRAH). Los análisis de RMSD se realizaron usando como referencia las coordenadas de la estructura inicial previa a la minimización.

Los contactos nativos ADN-proteína fueron calculados considerando los contactos globales y los específicos de la cadena lateral en el surco mayor. Para el primer caso (contactos globales ADN-proteína) se definió un contacto como cualquier *bead* proteico que

esté hasta a 8 Å de los *beads* centrales de las bases nitrogenadas (ver tabla 2.2, *beads* N1A, N3T, N3C, N1G). En tanto, para el segundo caso (contacto ADN-proteína específico) se definió que un contacto existe cuando un *bead* que represente a la cadena lateral esté hasta a 9 Å de los *beads* centrales de las bases de ADN. En ambos casos se reporta el porcentaje de conservación de contactos a lo largo de la dinámica así como también el valor de accuracy (Acc) definido como Acc=TP/(TP+FP) donde FP simboliza los contactos generados durante la dinámica (falsos positivos) y TP los contactos nativos conservados.

Los análisis de cluster de las trayectorias de leucine zippers se realizaron mediante la aplicación g_cluster de Gromacs escogiendo el algoritmo de gromacs con un criterio de 0.3 nm de RMSD para definir estructuras vecinas presentes en el mismo cluster. Este valor fue tomado como referencia dado que en proteínas que presentan dinámicas estables con SIRAH existen fluctuaciones entorno a 3 o 4 Å.

IV RESULTADOS

4.1 Flexibilidad del ADN en presencia de iones

En una etapa inicial del trabajo aquí propuesto se estudiaron propiedades dinámicas del ADN debido a interacciones de iones en el surco menor del ADN. Se demostró que la presencia de iones en el surco menor inducía a una reducción en el ancho del mismo, efecto también conocido como *narrowing*. El hecho de haber podido muestrear mediante técnicas de dinámica molecular simples (ensemble NPT) estructuras con *narrowing* teniendo como punto de partida un ADN doble hélice de tipo B, mostró la posibilidad de que el modelo existente de ADN pudiera representar regiones con *narrowing* así como *bending* inducido de una manera dinámica. A su vez, también se compararon las ocupaciones de los iones en el surco menor con los datos de densidad electrónica reportados en el PDB para dicha estructura. La hipótesis estudiada, es que algunas de las regiones de densidad electrónica en el surco menor podrían corresponder de hecho a iones de Na^+ cuyo patrón de difracción es muy similar al del agua (Shui, McFail-Isom et al. 1998; Dans, Darré et al. 2013) En figura 2 se resumen los resultados para el caso de una secuencia de ADN cuya estructura está resuelta por cristalografía que fue simulada por 12 μs con el mismo protocolo explicado en la metodología. En la figura 2C se observa que en la región con *narrowing* en el surco menor, la existencia de ocupaciones de iones de sodio (verde) y potasio (azul) resultan solapantes con aguas cristalográficas reportadas (esferas rojas). Para más detalles ver Anexo II.

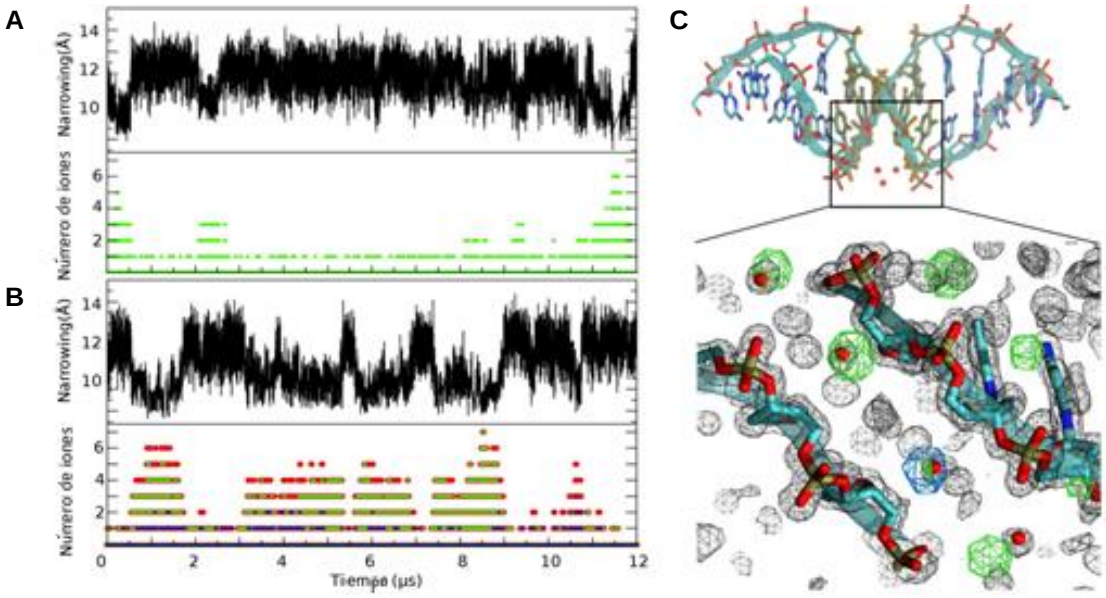


Figura 2: *Narrowing* inducido por iones en secuencia de dodecámero Drew-Dickerson de ADN. A y B muestran el perfil de *narrowing* a lo largo de una simulación de 12 μs junto con la presencia de iones en el surco menor para un sistema de ADN en presencia de iones de Na^+ (A) y en presencia de iones de Na^+ y K^+(B). En C se muestra un esquema de ocupación de iones en el surco menor contrastado con datos experimentales de cristalografía (PDBID 355D). Las grillas grises corresponden a la densidad electrónica reportada mientras que los colores azul y verde corresponden a iones de K^+ y Na^+ en A B y C. En B los puntos rojos corresponden a la suma de ambos tipos de iones.

4.2 Caracterización y análisis del campo de fuerza para proteínas

Como se explicó anteriormente, este campo de fuerzas está basado en la representación simplificada de proteínas mediante la reducción de la cantidad de átomos del sistema y el aumento del paso de tiempo.

Parte de la tesis consistió en el desarrollo y caracterización del campo de fuerzas de proteínas. Para cumplir con este objetivo fue necesario realizar varias simulaciones de dinámica molecular sobre sistemas que comprendieran un conjunto representativo de varias topologías de proteínas. Como una versión simplificada de elección de sistemas se decidió elegir estructuras que de acuerdo a la categorización de SCOP(Lo Conte, Ailey et al. 2000) fueran mayormente alfa, mayormente beta, alfa/beta o alfa+beta. También fue necesario elegir una serie de características a evaluar para examinar de manera sistemática la calidad de los resultados obtenidos.

Una vez obtenido un set de parámetros adecuados del campo de fuerza, se realizaron corridas de diferentes sistemas para validar el campo de fuerza de proteínas. En esta sección se presentan resultados provenientes de un sistema compuesto por la proteína top7(Kuhlman, Dantas et al. 2003) (PDB ID 1QYS) para la caracterización del campo de fuerza de proteínas. Esta proteína fue originalmente diseñada *in silico* y fue demostrada por tener un *folding* extremadamente estable. Debido a esto y a que es una proteína globular pequeña que presenta los elementos de estructura secundaria más representativos: alfa hélices (2) y hoja beta (5 hebras beta) se eligió para realizar la validación estructural. En la figura 3A se puede apreciar el *folding* de esta proteína así como los elementos de estructura

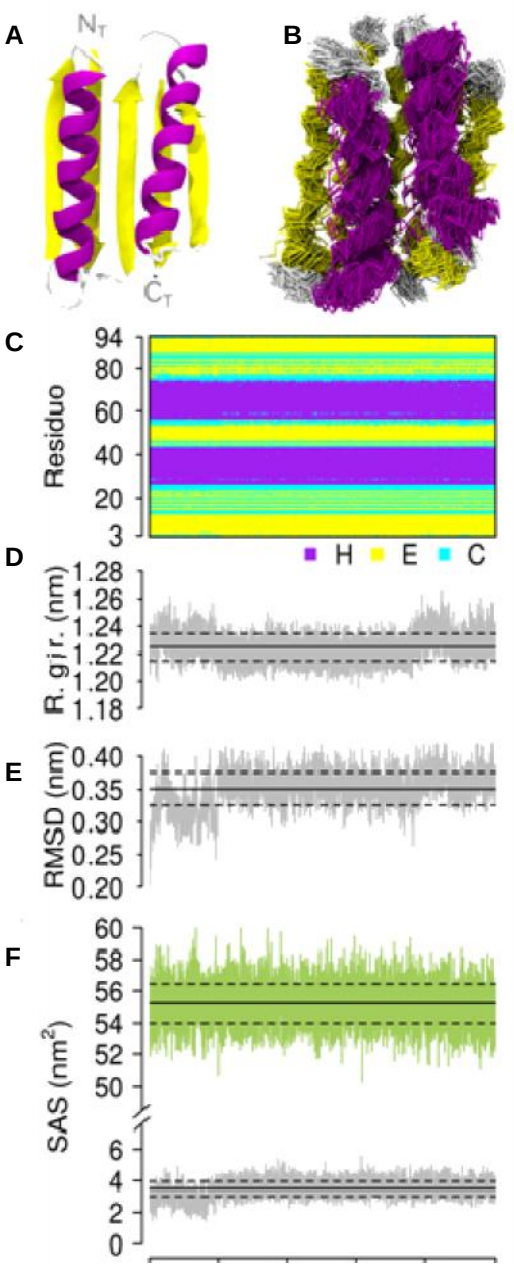


Figura 3: Simulación de proteína Top7.
Representación atomística de estructura de Top7 (A) y estructuras tomadas cada 10ns de dinámica coloreadas por elementos de estructura secundaria donde la hélice es púrpura, las hojas beta amarillas y el resto blanco(B). Estructura secundaria de cada residuo a lo largo de la dinámica(C). Radio de giro y RMSD calculados en función del tiempo sobre átomos GC (D, E). Superficie accesible al solvente hidrofílica (F, arriba) e hidrofóbica (F, abajo). En D-F las líneas horizontales enteras representan el promedio y las punteadas los desvíos estándar.

secundaria que la componen. El sistema simulado consistió en la proteína modelada con sus terminales neutros debido a que los primeros y últimos residuos de la misma no están reportados en la estructura reportada pero si estaban presentes en el cristal y se trabajó únicamente con los aminoácidos determinados estructuralmente. Se simuló entonces a la proteína en presencia de 1033 aguas *coarse-grained* WT4 y 30 NaCl, que representa una concentración salina aproximada a 150 mM. Una vez minimizado y equilibrado el sistema se corrió una dinámica de producción por 1 μs. En la figura 3B se muestra la superposición de las estructuras que adopta top7 durante la dinámica de producción tomadas cada 10ns. Aquí se puede observar que la estructura secundaria aparenta conservarse bien a lo largo de la dinámica como también la estructura terciaria. En 3C se puede ver que calculando la estructura secundaria de cada residuo en función del tiempo, ésta permanece extremadamente estable, con pequeñas fluctuaciones sobre los extremos de los elementos estructurados que puede ser interpretado como consecuencia de efectos de movimientos térmicos. Para caracterizar mejor la variación en la estructura global de la proteína se decidió calcular el radio de giro de la misma calculado sobre los *beads* GC (figura 3D). Éste se estabilizó oscilando en torno a 1.22 nm luego de 0.2 μs de dinámica un valor muy cercano al inicial (1.21 nm). Análogamente, el análisis de la desviación cuadrática media (RMSD) calculado para los GC a lo largo de la dinámica muestra que existencialmente un leve aumento y descenso en los primeros 0.15 μs para alcanzar finalmente un valor estable entorno a los 0.36 nm (figura 3E). También resulta interesante estudiar la superficie accesible al solvente de la proteína discriminando superficies hidrofóbicas e hidrofílicas para descartar casos donde cadenas laterales hidrofóbicas podrían haber rotado para quedar expuestas al solvente. En la figura 3F se puede observar que ambas superficies se mantienen estables a lo largo de la dinámica con un pequeño aumento de la superficie hidrofóbica en los primeros 200ns. Sin embargo, al promediar los valores antes y después de ese punto, los promedios varían únicamente 0.8 nm^2, lo que comparado con la superficie hidrofóbica de una alanina atomística que es 1.1 nm^2 representa una diferencia muy pequeña. Finalmente, se calcularon la conservación de contactos nativos globales, definidos como aquellos presentes en la estructura cristalográfica. Se definió que un contacto nativo global existe si dos átomos GC están a 8Å o menos entre sí en la estructura cristalográfica, mientras que un contacto nativo local además de cumplir este requisito debe ser entre dos residuos que estén separados por al menos cuatro posiciones en la secuencia primaria. Considerando esto, los contactos nativos locales se conservaron un 75% (s.d. 3) con accuracy 64% (s.d.3) mientras que los contactos globales alcanzaron valores de conservación de 89% (s.d.1) con accuracy de 84% (s.d. 1)

Otras descripciones detalladas de diversos sistemas de proteínas pueden verse en el Anexo II.

4.3. Caracterización y reparametrización de arginina y lisina con ADN

El desarrollo del campo de fuerza de proteínas como complemento al ya existente modelo para ADN motiva el interés de combinar los dos modelos. Dado que ambos campos de fuerzas siguen el mismo enfoque estructural de desarrollo y en particular ambos se encuentran implementados para su uso con el solvente *coarse-grained* junto con sus iones, se decidió testear el uso conjunto de los mismos.

Como primera estrategia para estudiar complejos ADN-proteína con este campo de fuerza, se probó simplemente de correr simulaciones con los parámetros existentes para ácidos nucleicos y aminoácidos. Esto generó estructuras muy distorsionadas, principalmente a causa de interacciones entre cadenas laterales catiónicas y fosfatos del ADN. También se observaron en algunos casos extremos, exposición de bases al solvente así como intercalación artificial de algunas cadenas laterales entre las bases de ADN, distorsionando tanto la estructura global como local del ADN. Por este motivo se decidió proseguir con la reparametrización de interacciones entre ambas especies macromoleculares.

Rohs y colaboradores estudiaron la frecuencia de aminoácidos en contacto con el surco menor del ADN, reportando que arginina y lisina son los aminoácidos con mayor frecuencia en el surco menor (30% y 10% respectivamente) y en particular en surcos menores con *narrowing* ese porcentaje se eleva hasta 60% y 15% respectivamente (figura 4). (Rohs, West et al. 2009[1-2]) Como el modelo CG de ADN demostró que puede representar estructuras con *narrowing* y puesto que la mayoría de las alteraciones provenían efectivamente de interacciones sobredimensionadas entre aminoácidos catiónicos y los fosfatos del ADN, se decidió realizar un primer paso de corrección involucrando inicialmente solamente las interacciones con estos residuos. Como sistema de control se consideró una doble hebra de ADN en presencia de solvente CG (WT4) donde la carga total del sistema fue neutralizada con cadenas laterales de arginina o de lisina CG. Esto facilitó un rápido muestreo de los sitios de interacción favorecidos y permitió elaborar perfiles de distribución de distancias entre átomos de lisina y ADN o de arginina y ADN respectivamente. Como referencia experimental de perfil de distribución de distancias representativo de todas las estructuras de complejos ADN-proteína se utilizaron los datos

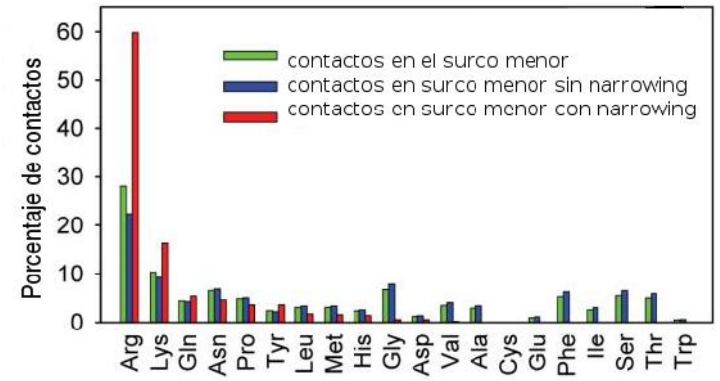


Figura 4: Frecuencia de aminoácidos en el surco menor. En verde se señalan los porcentajes de aminoácidos encontrados en el surco menor. En azul y rojo se representan los porcentajes de contactos en el surco menor en ausencia o presencia de *narrowing* respectivamente. Adaptado de Rohs et al(Rohs, West et al. 2009[2]).

provenientes de la base de datos de interfases PDIdb (***P**rotein **D**NA **Interface d**ata**b**ase*)(Norambuena and Melo 2010). Esta parte fue realizada en conjunto con el grupo del Dr. Francisco Melo, cuya especialidad es el estudio de interacciones proteína ADN por métodos bioinformáticos. En particular la experiencia del mismo en elaborar análisis estadísticos representativos del universo de dichas interacciones permitió establecer una estrategia para el estudio comparativo de los resultados CG y experimentales. PDIdb tiene la ventaja de ser una base de datos que está curada y es un set no redundante representativo de todos los complejos reportados en el PDB Para el análisis de los datos experimentales aquí presentados se trabajó en conjunto con el fin de generar histogramas de

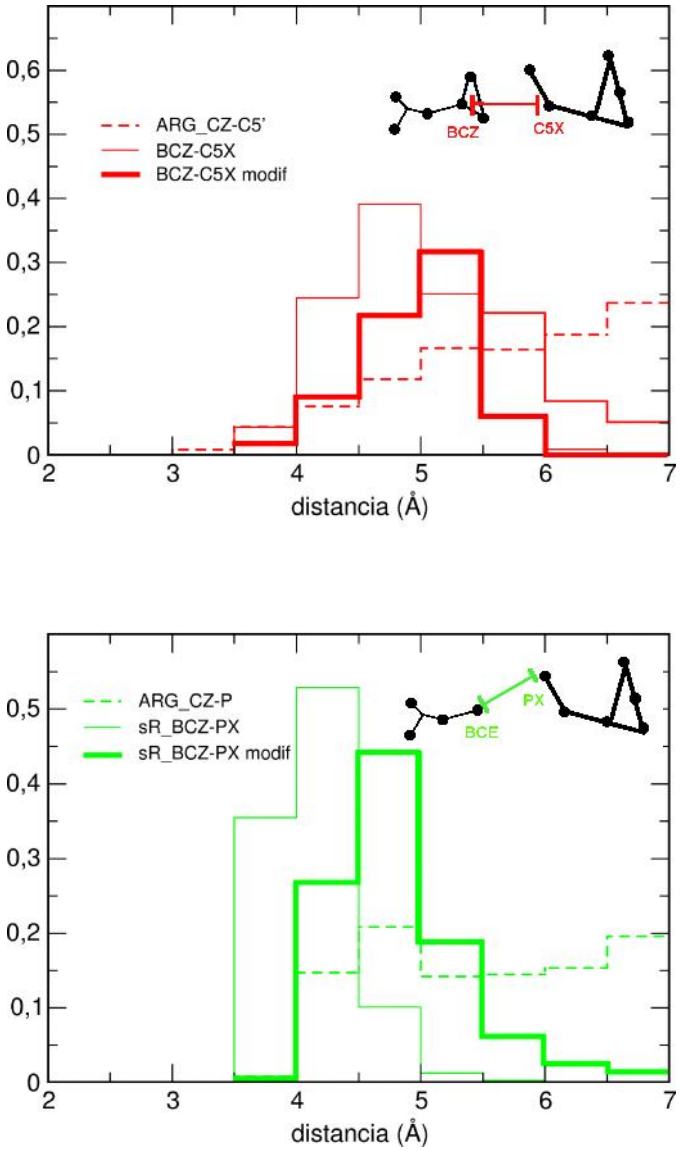


Figura 5: Perfiles de distancia entre átomos seleccionados de arginina (CZ) y ADN (C5',P). En A y B las líneas con trazo cortado indican valores estadísticos obtenidos desde datos globales de PDIdb, las enteras finas indican valores resultantes de una simulación sin modificaciones y las más gruesas reportan las distancias resultantes de simulaciones con modificaciones en los parámetros de interacción BCZ-PX

distancia de pares de átomos de interés que sean representativos del universo de complejos ADN-proteína. Para ello se generaron una base de datos de interfases no redundantes desde estructuras depositadas en el PDIdb, considerándose únicamente aquellos complejos cristalográficos que tengan una resolución de hasta 2.5 Å para calcular los histogramas de pares de átomos de ADN con lisina o arginina con valores de hasta 7Å. Dado que el interés está en representar correctamente la interfase entre moléculas de ADN y proteína, se optó por analizar solamente aquellas interacciones presentes en una distancia de hasta 7 Å entre estos átomos.

Al comparar estos perfiles de distribución con los datos obtenidos de estructuras de complejos ADN-proteína curadas existentes en el PDIdb se pudo identificar claramente que existía una mayor afinidad entre los pares de átomos BCE-PX y BCZ-PX, ya que la mayoría de

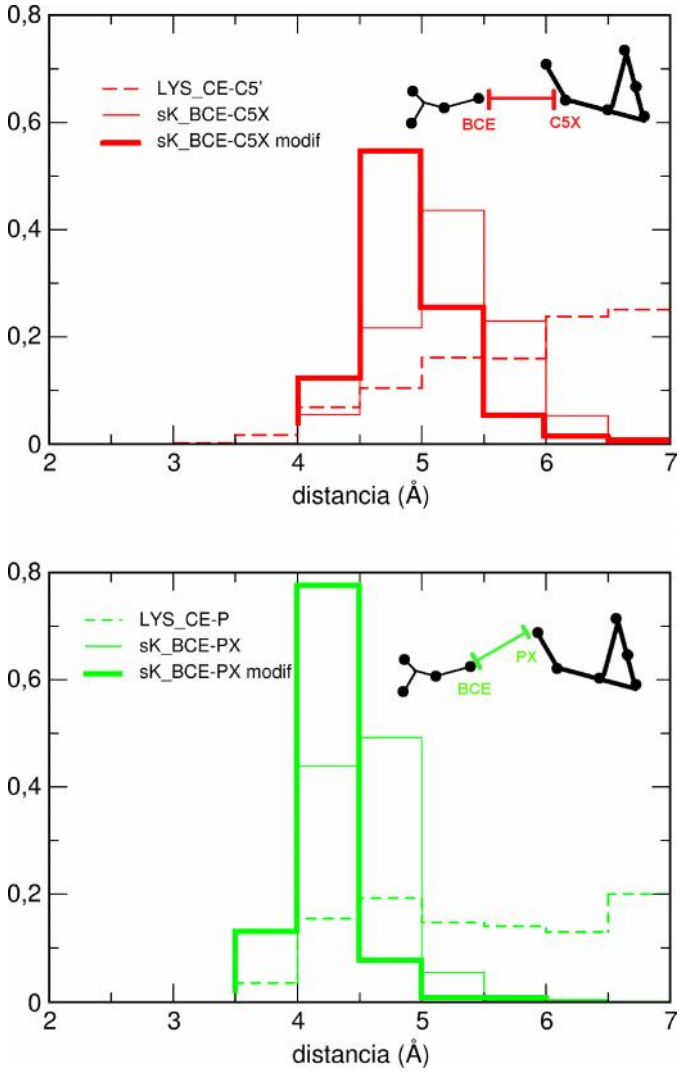


Figura 6: Histogramas de distancias entre átomos CE de lisina y P, C5' de ADN. Las líneas rayadas indican valores estadísticos obtenidos desde datos de PDIdb, las enteras indican valores resultantes de una simulación inicial sin modificaciones y las punteadas reportan las distancias resultantes de simulaciones con modificaciones en los parámetros de interacción BCE-PX

las interacciones se reportaba a menores distancias y con mayor predominancia. En la figura 5 se representa el histograma de las distancias entre arginina y fosfatos de los valores experimentales (línea a trazos, ARG_CZ_P) y de los valores iniciales obtenidos de la dinámica de ADN con cadenas laterales de arginina (sR_BCZ-PX). Se puede observar que el primer pico correspondiente a sR_BCZ-PX se encuentra entre 3,5 y 4 Å mientras que los valores experimentales (ARG_CZ-P) lo reportan entre 4 y 4,5 Å teniendo el máximo entre 4,5 y 5 Å.

Por otro lado, en la figura 6 que representa los histogramas análogos para lisina, se puede observar que el primer pico para la distancia entre el carbono ε y el fósforo del fosfato se encuentra entre 4 y 4,5 Å mientras que el máximo se encuentra entre 4,5 y 5 Å. Aunque los primeros picos se observan en las mismas regiones que los datos experimentales (línea a trazos, LYS_CE-P), se encuentran sobredimensionados entre los valores 4-5Å en comparación con los experimentales que se encuentran más distribuidos entre 4 y 7 Å.

Como se desea generar un perfil similar al observado en el PDIdb de manera de no perder interacciones características debido a la simplificación del modelo, se decidió entonces aplicar modificaciones a las interacciones específicas entre átomos BCE y PX y BCZ y PX. Esto se realizó de manera de definir explícitamente los valores de σ y ε para la interacción de específica tipo Lennard-Jones entre esos átomos. Se consideró esta opción dado que no se deseaban cambiar los tipos de átomos sino únicamente la interacción no enlazante entre estas entidades. Esto es posible de realizar en GROMACS, donde está implementado SIRAH proteínas, ADN y solvente. Para definir un primer *guess* del valor de la interacción entre dos átomos A y B es necesario definir explícitamente los valores para $\sigma_{A,B}$ y $\varepsilon_{A,B}$ utilizados en la potencial no enlazante de tipo Lennard-Jones (L-J) (ver tabla 2.1 en fundamentación). Teniendo en cuenta que el mínimo del potencial de L-J en GROMACS no corresponde con el σ de la interacción (que es el valor en donde el potencial vale 0) sino con la distancia de equilibrio, se estimó el primer *guess* como $\sigma_{int=} r_{min}/2^{1/6}$ donde r_{min} corresponde al valor observado del pico máximo en las distribuciones del histograma generado a partir de datos experimentales. Este proceso se realizó de manera iterativa alterando los valores de σ para representar correctamente el primer pico de distribución del par de átomos. En la tabla 4.1 se muestran los valores de σ y ε iniciales sin modificar y los finales. Los valores de ε fueron modificados para regular la afinidad d ella interacción, y por ende este cambio se ve reflejado en la forma de la distribución.

Tabla 4.1: Valores iniciales y finales de σ y ε para las interacciones entre pares de átomos modificados mostrados en los perfiles de distancia.

	PX-BCE (sK)	PX-BCZ (sR)
$(\sigma_{int;}\ \varepsilon_{int})$	0.50; 0.68	0.43; 0.48
$(\sigma_{int;}\ \varepsilon_{int})$modificado	0.55; 0.30	0.54; 0.1

Los valores obtenidos luego de la modificación corresponden a las líneas gruesas (sR_BCZ-PX modif, sK_BCE-PX modif) en las figuras 5 y 6. En el caso de la arginina tiene su primer pico entre 4 y 4,5 Å y el máximo entre 4,5 y 5 Å como los datos experimentales. Si bien, es posible no haber logrado un perfil igual de manera cuantitativa al experimental, es posible afirmar que a nivel cualitativo se tiene un perfil similar a cortas distancias, que es razonable dado que se trabajó únicamente con cadenas laterales en solución y no estructuras proteicas completas. Esto tiene como consecuencia que no se estarían perdiendo interacciones cercanas, o en este caso particular no estarían siendo sobredimensionadas como se observaba con anterioridad. Por otro lado, en el caso de la lisina los valores obtenidos luego de la modificación revelan un primer pico entre 4 y 4,5 Å y el máximo entre 4,5 y 5 Å como los datos experimentales.

Debido a las limitaciones de este primer sistema minimalista para la caracterización de las distancias de interacción entre ADN y lisina o arginina se decidió probar estas modificaciones sobre sistemas con ADN en presencia de proteínas completas.

4.4 Caracterización de arginina y lisina en complejos ADN/leucine zipper

Los parámetros modificados para interacciones de lisina y arginina con fosfatos fueron agregados al campo de fuerza y probados consecutivamente con un set minimalista de complejos conteniendo proteínas de tipo leucine zipper. Este conjunto no redundante de complejos con leucine zippers fue escogido debido a la relativa estabilidad que tienen estos complejos unas vez dimerizados, minimizando de esta forma las posibles distorsiones debidas a las variaciones estructurales intrínsecas de la proteína a lo largo de la simulación. En estos sistemas se agregó además una modificación extra para la interacción entre el *backbone* de la proteína y las bases nitrogenadas de manera que estos átomos que, entre su misma especie interactúan con parámetros que los hacen ser prácticamente atomísticos, también interaccionen de manera análoga entre los dos tipos distintos de moléculas. Este cambio no fue implementado en el caso anterior dado que se simulaba únicamente la cadena lateral de los aminoácidos de interés.

A modo de ejemplo se detallará seguidamente uno de los sistemas simulados. La estructura se eligió por presentar un motivo sencillo de tipo leucine zipper perteneciendo a la familia de proteínas bZIP y superfamilia según SCOP de leucine zippers y se eligió también por presentar un modelo sencillo de leucine zipper con el dominio básico de unión a ADN y el dominio zipper, la hélice alfa

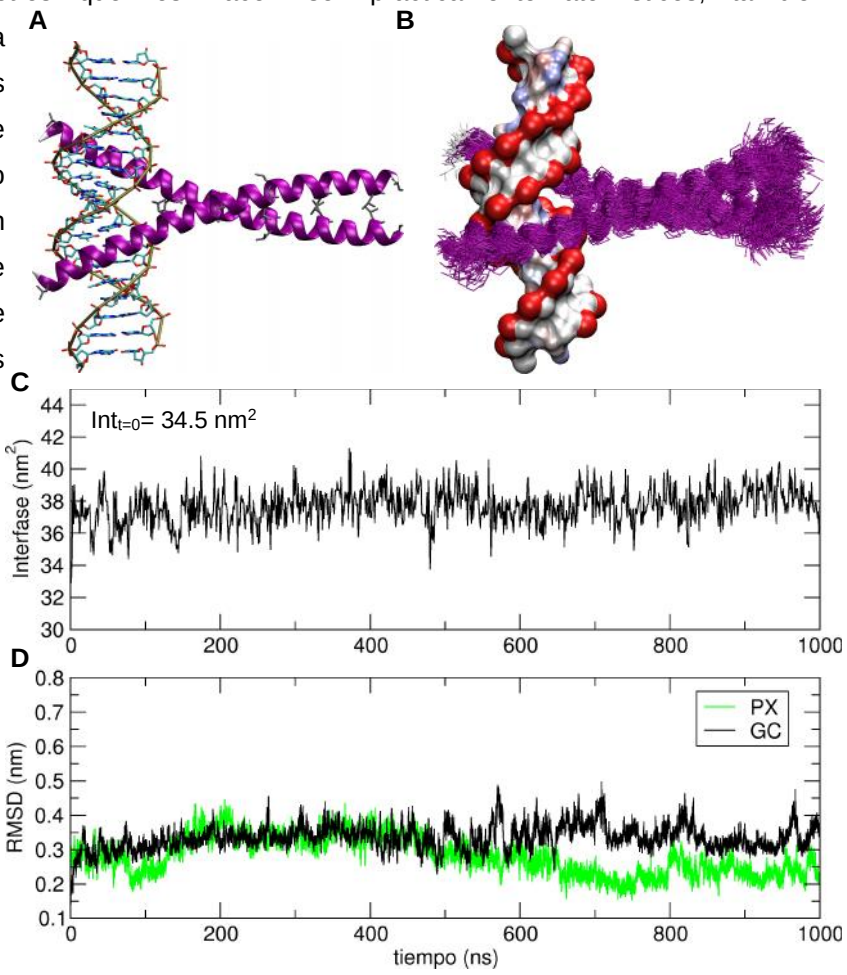


Figura 7: GCN4-bzip/ATF/CREB. En A y B se presenta la estructura del complejo atomística y CG coloreada por estructura secundaria de la proteína (alfa hélice: morado). En B se representa la superficie accesible al solvente del ADN coloreada por carga (rojo: negativo, azul:positivo) junto con estructuras proteicas tomadas cada 1 ns durante la dinámica del sistema coloreadas también por estructura secundaria. Se muestra la evolución de la interfase junto al valor inicial correspondiente a la estructura cristalográfica (C) y perfiles de RMSD calculados sobre todos los *beads* PX y sobre los GC correspondientes a regiones con estructura secundaria asignada (D).

que permite la dimerización con otra análoga a través de repeticiones de leucinas cada 7 aminoácidos aproximadamente. La estructura se encuentra resuelta por cristalografía con una resolución de 2,20 Å y está compuesta por el dímero GCN4-bzip y una secuencia palindrómica de ADN que contiene el sitio de reconocimiento de ATF/CREB (secuencia ATGAC/GTCAT) (figura 7 A)(Keller, König et al. 1995). Los extremos C- terminal de la proteína forman el leucine zipper que se encuentran orientados perpendiculares al eje de simetría del ADN, mientras que el dominio N-terminal de cada monómero se encuentra conformado por la región básica de unión al ADN, y por lo tanto está formando contacto con el surco mayor en una región comprendida por 12 pares de base que incluyen el motivo de reconocimiento. La interfase está compuesta tanto por interacciones a nivel del surco mayor como también por interacciones a nivel del *backbone* de fosfatos. Algunas interacciones específicas conservadas suceden entre los aminoácidos sN235, sR243 y sK246.

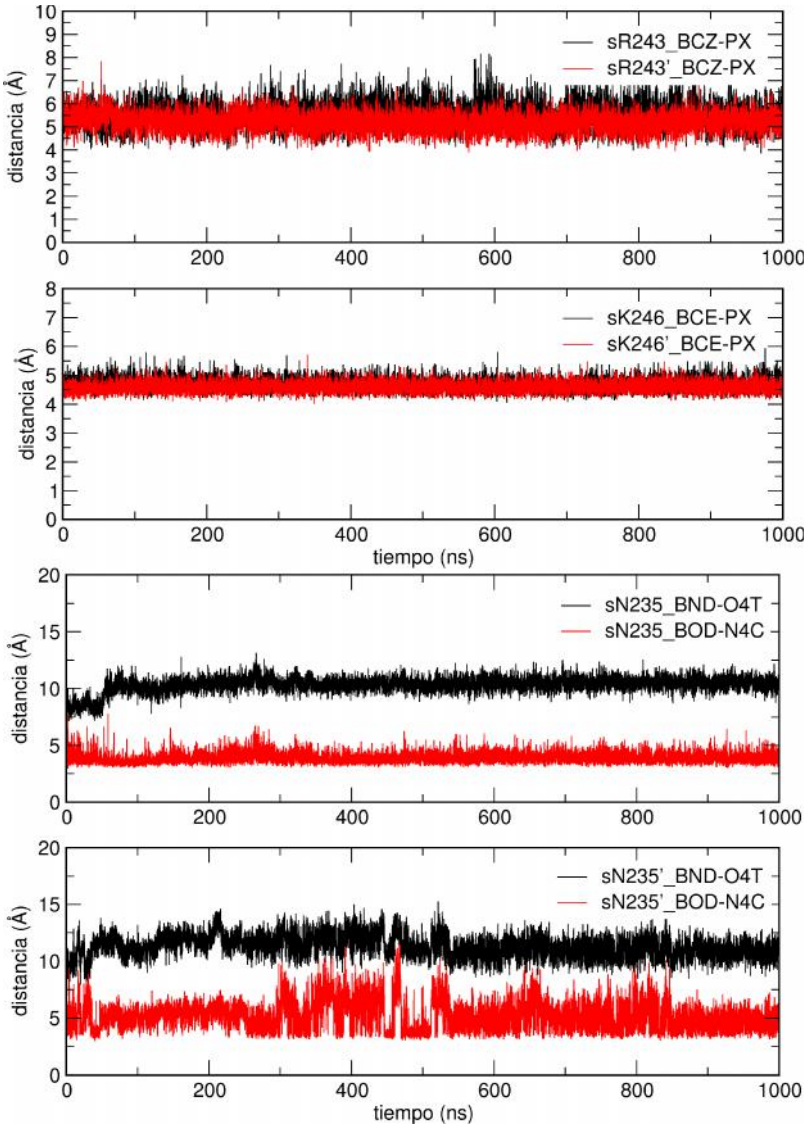


Figura 8: Distancias específicas entre interacciones características. Distancias entre átomos representativos de interacciones conservadas en GCN4-bzip/ATF/CREB

Con el fin de obtener un análisis más detallado se estudiaron las distancias entre los átomos envueltos en las interacciones anteriormente mencionadas a lo largo de la simulación tal como se puede observar en los gráficos de la figura 8. sN235, que está dirigida hacia el surco menor estableciendo interacciones con las posiciones 3 y 4 de la secuencia nucleotídica, presenta perfiles estables a lo largo de la simulación luego de los primeros 50 ns alcanzando valores

similares para los mismos aminoácidos de cada hélice donde sólo es visible una mayor fluctuación en el caso de sN235 de la primera hélice. Por otro lado, sR243 que además de estar interaccionando con G1 en el surco mayor lo hacen con el fosfato -1 se presenta extremadamente estable a lo largo de toda la dinámica para ambas subunidades del dímero. El tercer aminoácido conservado reportado por los autores, sK246 se encuentra presente en el límite del dominio de unión básico al ADN y la región propiamente dicha del leucine zipper, éste interacciona principalmente a través de su cadena lateral en el surco mayor con las bases en posición -3. En cristalografía esta interacción está propuesta por encontrarse mediada por agua y el átomo Cε se encuentra a una distancia de de entre 4.3 y 4.7 Å aproximadamente del fosfato correspondiente a la base con la que establece enlaces de hidrógeno mediados por agua. Para el caso de la simulación *coarse-grained* se puede apreciar que los valores representativos de estas interacciones con lisina se mantienen estables a lo largo de la dinámica entorno a valores de entre 4.5 y 5 Å, que coinciden con los presentes en la cristalografía.

Asimismo, se realizó un análisis global de la estabilidad del sistema considerando como parámetros el área de la interfase así como el perfil de RMSD representativo tanto de la proteína como del ADN. Para representar la estabilidad de la proteína se consideró el cálculo de RMSD sobre los átomos GC (equivalentes al Cα) que presentan una estructura secundaria determinada, mientras que para analizar la estabilidad del ADN se lo hizo sobre los átomos correspondientes a los fosfatos del *backbone* (PX). En la figura 7C se puede apreciar que la interfase del sistema tiene oscilaciones de alrededor de 4 nm2 en los primeros 200 ns para luego estabilizarse en el entorno de los 38 nm^2 donde se aprecia un leve aumento respecto al valor a t=0 (previo a la minimización) que corresponde a 34.5 nm^2. El RMSD para la proteína se estabiliza rápidamente entorno a un valor de 0.35 nm, mientras que el RMSD calculado sobre los fosfatos presenta un perfil con mayores oscilaciones a lo largo de la dinámica si bien también se estabiliza entorno a valores bajos de 0.3 nm aunque presenta un leve descenso en los últimos 300 ns alcanzando finalmente en valores alrededor de 0.25 nm. Esto se debe a la estabilización final del extremo 5' del ADN cuyas bases se encuentran desfasadas al principio de la dinámica (*staggered*) a causa de la presencia de *narrowing* en el surco menor. Considerando los valores presentados de variabilidad estructural en una proteína estable como la top7 presentada en la sección 4.2, que para el caso de RMSD sobre los GC reportaba valores en torno a 0.36 nm podemos afirmar que este complejo GCN4-bzip/ATF/CREB presenta valores de RMSD estables considerando que se trata de un sistema *coarse-grained*. Dicha conclusión también se encuentra soportada por el análisis de las distancias de interacciones específicas de este sistema tales como fueron mostradas en la figura 8. Estos resultados, junto con la conservación a lo largo de la dinámica de los contactos globales nativos (46% con accuracy asociada de 53%) y la de conservación de contactos de cadenas laterales (54% con accuracy de 50%) servirán como referencia para valores de un complejo

estable ADN-proteína.

A continuación se presentan los perfiles de distancias entre átomos de interés de lisina y arginina para todo el subset completo de leucine zippers estudiado. Es importante resaltar que la comparación del conjunto simulado se realizó únicamente con una base de datos creada a partir de los datos cristalográficos de las mismas estructuras que fueron simuladas a diferencia del caso anterior (sección 4.4) que se comparaba contra datos estructurales representativos de todo el universo de estructuras de complejos ADN-proteína reportados en el PDB hasta el 2010.

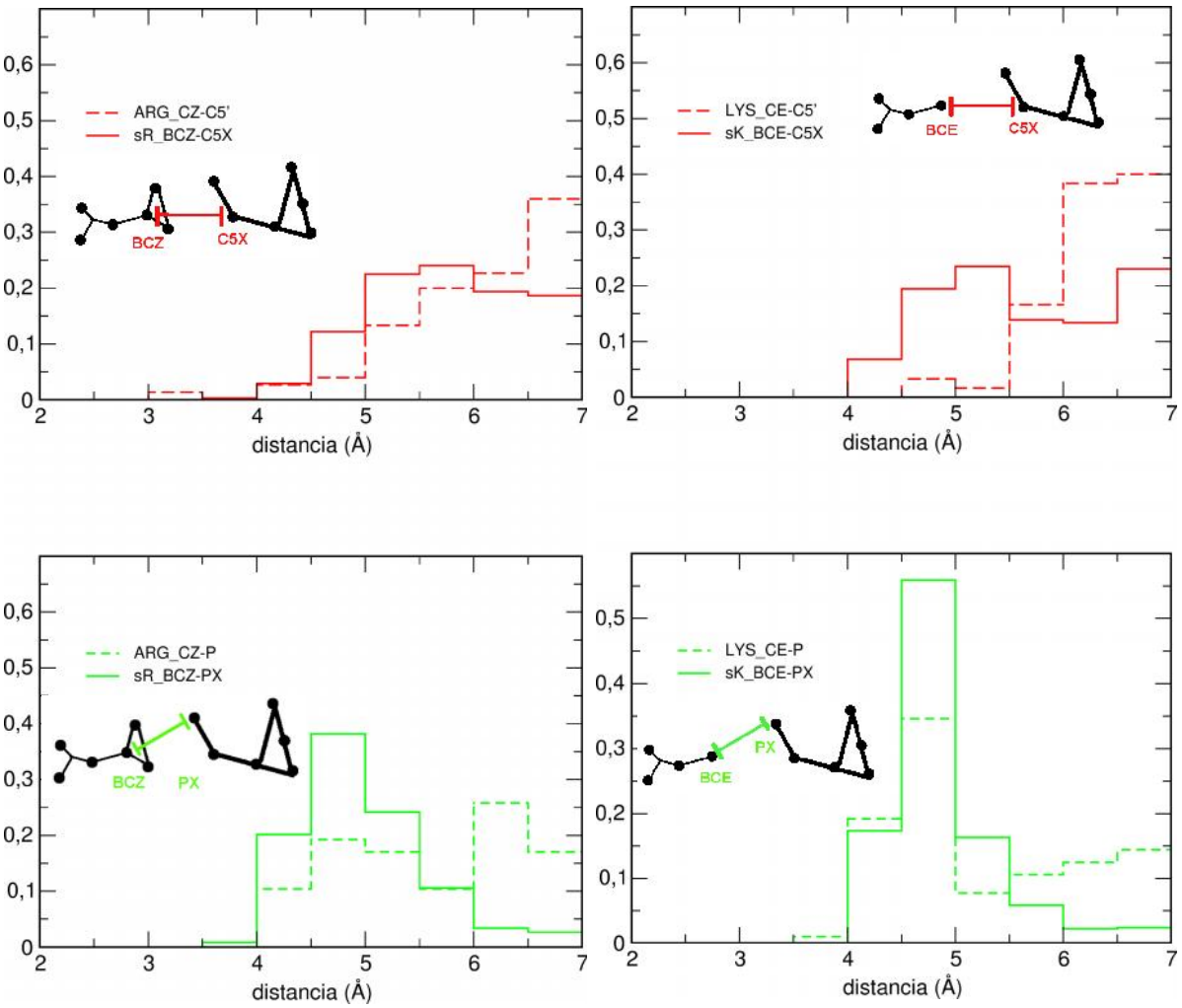


Figura 9: Distribución de distancias entre pares de átomos seleccionados de arginina y ADN (izq) y lisinia y ADN (der). Los datos corresponden únicamente a las estructuras pertenecientes al set de leucine zippers. Las líneas a trazas representan valores provenientes del PDIdb para dicho subconjunto, mientras que las líneas enteras corresponden a los datos provenientes del análisis de cluster de las dinámicas de todas las estructuras de dicho conjunto.

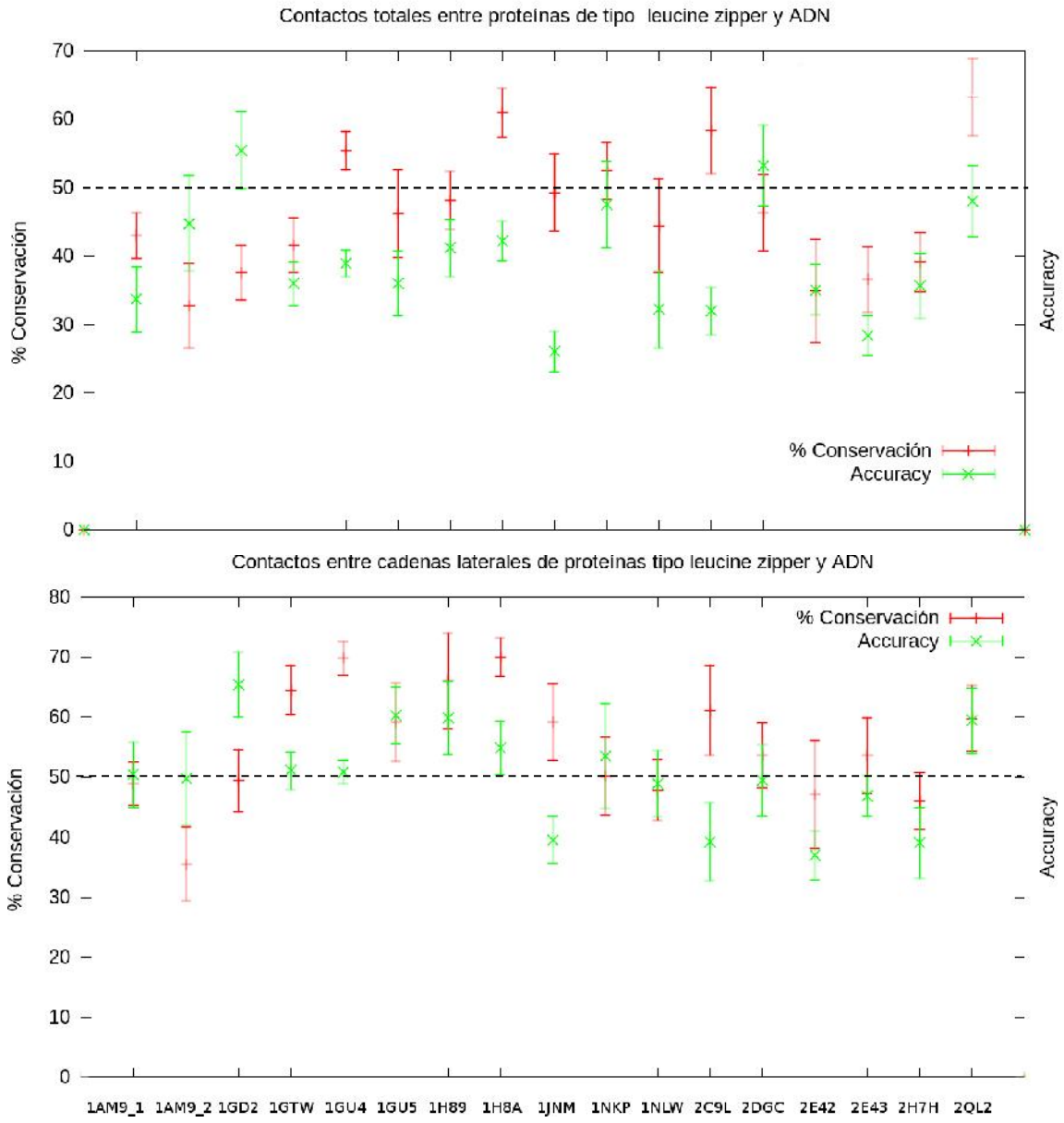


Figura 10: Conservación y accuracy de contactos para complejos ADN-proteína de tipo leucine zipper. Arriba se muestran los contactos globales y abajo aquellos que suceden con la cadena lateral únicamente. En ambos casos se representan los valores promedio a lo largo de la dinámica de 1 μs con sus respectivos desvíos estándar. La línea punteada señala el 50%. Los sistemas están anotados con el PDB ID.

En la figura 9 se muestran las distribuciones de distancias correspondientes a los pares de interacciones entre el carbono ζ de la arginina o el ε de la lisina y el C5' y P de cualquier nucleótido (líneas a rayas), así como sus contrapartes CG (líneas completas). Aquí se puede observar que tanto en el perfil de lisina como arginina las interacciones con fosfatos se encuentran más localizadas sobre la región comprendida entre 4 y 5,5 Å generando que a mayores distancias los picos disminuyan en comparación con los valores experimentales atómicos. En el perfil de arginina se muestra que para las interacciones con el fosfato el primer

pico se encuentra entre 4 y 4,5 Å mientras que el máximo está ubicado entre 4,5 y 5 Å. Estos valores están cualitativamente de acuerdo con los experimentales. Análogamente, en el perfil de lisina es posible observar que el pico se encuentra entre 4 y 4,5 Å mientras que el máximo está ubicado entre 4,5 y 5 Å al igual que los datos experimentales. En este caso en particular es interesante observar la diferencia respecto al perfil observado inicialmente en la parte anterior donde sólo se simularon las cadenas laterales de los aminoácidos.

Para estos sistemas de leucine zippers además de mirar las distribuciones de distancias de los pares de átomos de interés, también se analizaron diferentes características propias de la interfase que sirvan para describirla cuali y cuantitativamente. Para ello se calcularon valores de conservación de contactos nativos tanto a nivel global como local (de cadenas laterales con el ADN, ver capítulo III Detalles computacionales). En la figura 10 se muestran los valores de conservación global (arriba) y local (abajo) de contactos junto con sus respectivas desviaciones estándar para cada uno de los complejos con leucine zipper. Además, en verde se muestran los valores de accuracy también con sus desviaciones estándar asociadas. El valor de accuracy complementa de manera cualitativa la calidad del modelo, ya que considera intrínsecamente la presencia de contactos artificiales generados durante la dinámica y no únicamente aquellos perdidos durante la dinámica.

Es decir, que un accuracy de 100% representaría una dinámica donde no se hubiera generado ningún contacto nuevo y se hubiera mantenido algún contacto original. Valores de accuracy mayores a 50% representan que dentro de la proporción de contactos existentes se conservaron más de los que se generaron, mientras que valores menores a 50% representan lo opuesto. En la figura 10 se marcan con líneas punteadas el valor de 50% para tener como referencia. Aquí se puede observar que al considerar la conservación local de contactos se obtienen mejores valores así como también de accuracy en comparación con la conservación global de contactos, donde la media para los valores de conservación y accuracy supera el 50 %. En Darré et al. reportamos que en algunos casos de complejos proteicos puede existir pérdida de contactos en la interfase. Esto sucede en el complejo trimérico formado por el dímero de HP1 Chromo-Shadow(CS) y el péptido CAF1. Si bien la superficie accesible al solvente total del sistema se conserva de manera estable (junto con su balance hidrofóbico/hidrofílico), en la interfase entre los dos monómeros del dímero CS los valores de conservación reportados son de 55% con accuracy de 26%. Esto indica un cierto grado de promiscuidad en los contactos debido a la granularidad y diminución de puntos de interacción del modelo. De cualquier modo, los contactos globales se conservaron un 80% con accuracy asociada de 76% y la interfase específica entre CAF1 y el dímero CS se mantiene estable. (Darré, Machado et al. 2015).

Una vez terminada esta etapa se decidió continuar con el análisis de sistemas más complejos, realizándose simulaciones de dinámica molecular de diversos complejos ADN-proteína con diferentes características estructurales.

4.5 Análisis de diversos complejos ADN/proteína

Se eligieron distintos tipos de complejos ADN-proteína con estructuras depositadas en el PDB y cuyas proteínas representaran diversos motivos de unión al ADN. Para esto se buscaron complejos que tuvieran interacciones también a nivel del surco menor, así como distintos motivos de unión típicos con el ADN que no fueron vistos en la parte anterior. En 4.4 se observó que los complejos que presentan interacciones con el ADN de proteínas de tipo leucine zipper son estables. Esto es esperable ya que estos sistemas no presentan grandes inconvenientes a la hora de ser modelados, puesto que el ADN está mayormente en conformación B y las proteínas son esencialmente alfa hélices. Por este motivo se decidió elegir para los sistemas de este apartado un conjunto de estructuras que presenten motivos más exigentes para el campo de fuerza y estudiar en función de ellos la viabilidad del mismo.

Con estos criterios se eligieron 10 complejos donde se encuentran los siguientes tipos de proteínas: proteínas centroméricas (PDB ID 1HLV)(Tanaka, Nureki et al. 2001), proteínas de *bending* de ADN (Factor de integración al hospedero-IHF, PDB ID 1OUZ)(Lynch, Read et al. 2003), proteínas de empaquetamiento del ADN (Histonas en nucleosoma, PDB ID 1KX5)(Davey, Sargent et al. 2002), proteínas de manutención y protección del ADN (PDB ID 2ZX3)(Shirai, Watanabe et al. 2009), proteínas de unión a TATA-box (TBP, PDB ID 1CDW)(Nikolov, Chen et al. 1996), proteína cromosomal hipertermófila Sac7d (PDB ID 1AZP)(Robinson, Gao et al. 1998), proteína de unión a ADN dependiente de secuencia (Dominio Homeobox PDB ID 2LKX, 1FJL)(Wilson, Guenther et al. ; Chaney, Clark-Baldwin et al. 2005), proteína de unión a ADN que presenta *bending* (proteína reguladora E2 del virus de papiloma humano, PDB ID 2BOP)(Hegde, Grossman et al. 1992) y finalmente un complejo multimérico compuesto por Ets-1/PAX5/ADN (PDB ID 1MDM)(Garvie, Pufall et al. 2002). Además, desde un punto de vista estructural, estos sistemas representan diversos modos de unión al ADN. A continuación se presentan los resultados detallados para tres de estos sistemas: estructura del core del nucleosoma, proteína de unión a TATA-box y factor de integración al hospedero-IHF. En el anexo I se presenta una tabla con valores de conservación de contactos para cada sistema simulado.

4.5.1 Estructura del core del nucleosoma

La estructura aquí estudiada (PDB ID 1KX5) corresponde al core de un nucleosoma resuelto a 1.9 Å de resolución. La molécula de ADN está compuesta por 147 pares de bases y proviene de la secuencia humana de un α-satélite que realiza 1.67 vueltas entorno al núcleo de histonas que corresponden a proteínas recombinantes de *X. laevis*. Esta estructura está compuesta por las histonas H2A.1, H2B.2, que conforman 2 dímeros (2 H2A-H2B) y las histonas H3 y H4 que se encuentran formando un tetrámero ($H3_2$-$H4_2$). Esta estructura se escogió en particular a causa de las interacciones de argininas en el surco menor, donde es

posible observar eventos de *narrowing*, de manera análoga a lo estudiado anteriormente con el ADN en presencia de iones. Éstas favorecen a estabilizar el *bending* del ADN y así mantener la estabilidad del complejo. En este sistema en particular se siguió el mismo protocolo explicado en materiales y métodos con la salvedad de agregar más pasos de equilibración gradual aumentando a 500ns la etapa previa de equilibración (datos no mostrados).

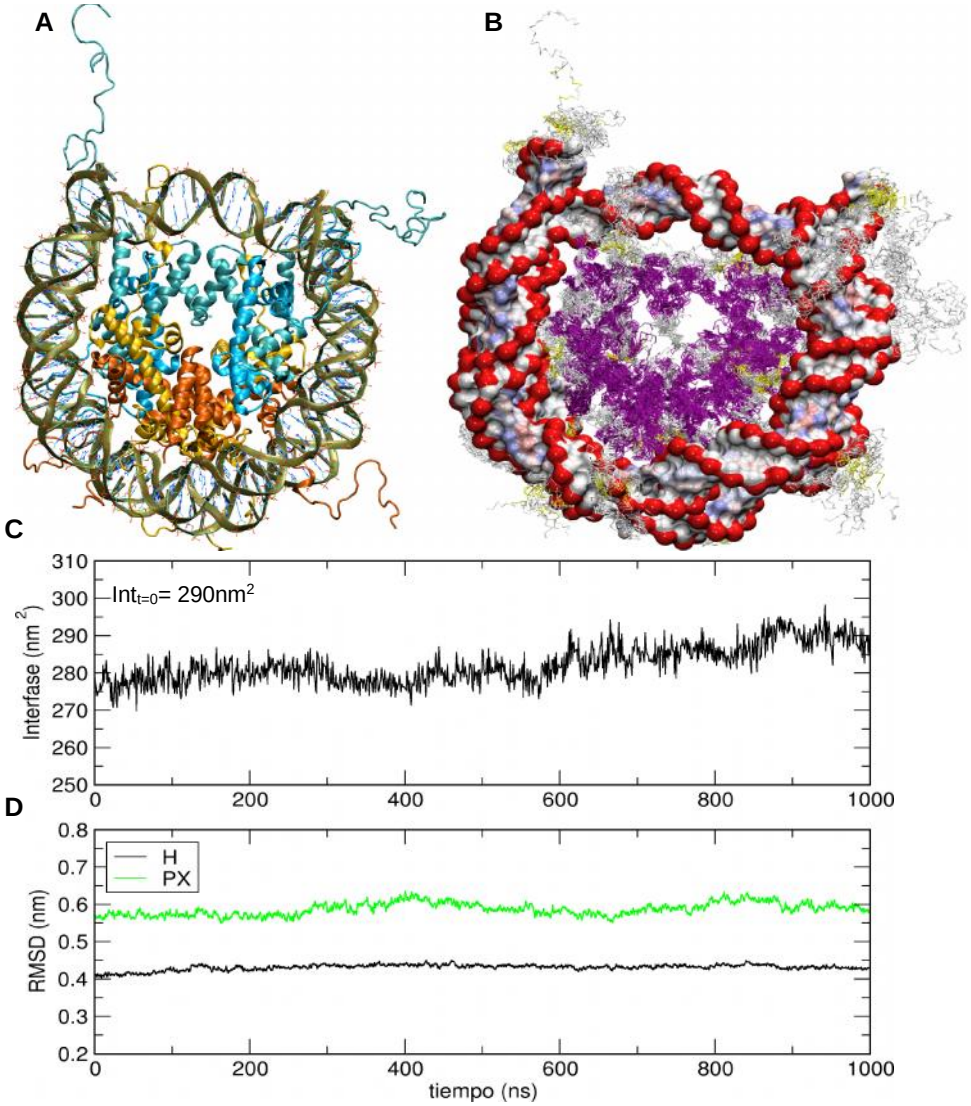


Figura 11: Estructura del core de un nucleosoma reportado en el PDB (A) y a lo largo de la simulación CG(B). En A las cadenas correspondientes a las diferentes histonas están representadas por color: H3 (turquesa), H4 (celeste), H2A (amarillo) y H2B (naranja). En B, el ADN está representado con su superficie accesible al solvente y coloreado por carga: negativo (rojo)-positivo (azul. Se puede observar que existe un ligero desenvolvimiento del ADN respecto al núcleo de histonas en los extremos, que es compatible con resultados reportados de dinámicas moleculares atómicas (Biswas, Langowski et al. 2013) Las proteínas están coloreadas por estructura secundaria: alfa hélice (violeta), hoja beta (amarillo) y se representan varias estructuras simultáneas de la proteína tomadas de la dinámica cada 100ns, la escala es levemente menor que en A para facilitar la visualización. En C y D se grafican los valores de la superficie de interfase y RMSD en función del tiempo de la simulación. Para facilitar la visualización se muestran los valores de RMSD calculados sobre todos los *beads* PX (verde) y sobre los GC que forman parte de estructuras de alfa hélice (negro) como promedios acumulativos calculados cada nanosegundo.

El perfil de la interfase a lo largo de la dinámica de producción (figura 11C) se mantiene relativamente estable, con un incremento a valores finales en torno a 290 nm^2. Un análisis más detallado de la interfase de la proteína en contacto con el ADN se obtuvo calculando de manera análoga a la anterior pero sin considerar los terminales flexibles de las proteínas que inicialmente no están formando contacto con la proteína. En la figura 12 se grafica la interfase del complejo depurando los terminales y normalizado por el valor inicial derivado de la estructura cristalográfica donde se puede apreciar que la superficie de la interfase se mantiene conservada en el orden del 90 %.

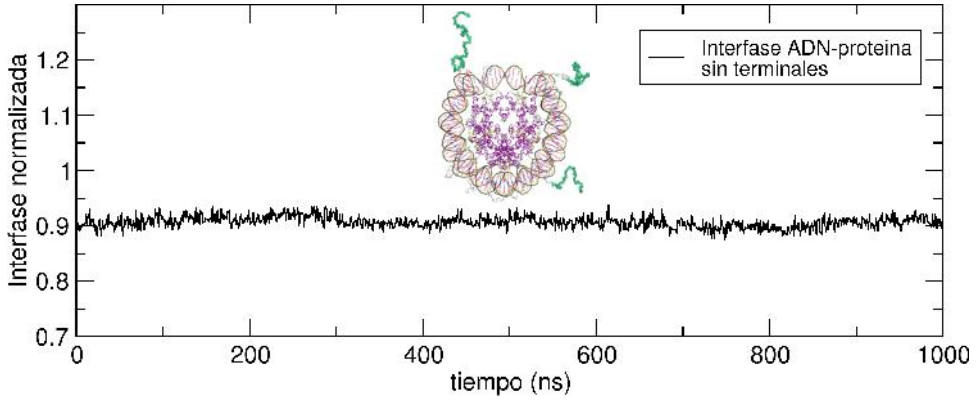


Figura 12: Interfase entre proteínas del core de nucleosoma y el ADN NCP147 calculada sin considerar los extremos N-terminales de las histonas que inicialmente no establecen contactos con el ADN (regiones del inset marcadas en verde). Los valores graficados corresponden a la interfase normalizada sobre el valor inicial derivado de la estructura cristalográfica.

Por otro lado el RMSD calculado sobre todos los fosfatos del ADN también se mantiene estable a lo largo de la dinámica oscilando entorno a los 5.5 o 6 Å. Para analizar la estabilidad de las histonas presentes se decidió calcular el RMSD sobre las regiones de proteína que tienen estructura secundaria de alfa hélice. Este motivo estructural es predominante en cualquiera de las

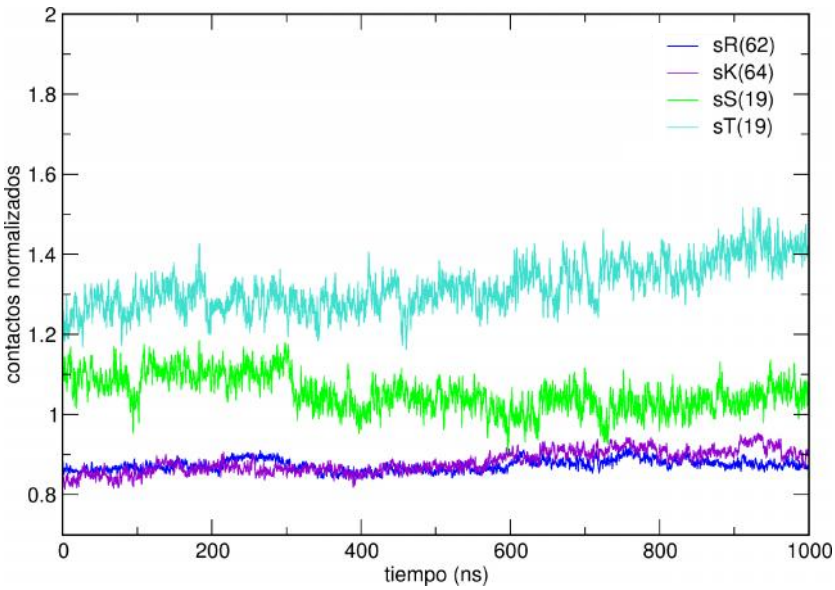


Figura 13: Contactos presentes de aminoácidos mayoritarios en la interfase histonas-ADN. Los valores están normalizados en cada una de las curvas por el valor inicial de contactos existentes para ese aminoácido en la estructura cristalográfica (valor que figura entre paréntesis en la leyenda) y graficados en función del tiempo de la dinámica de producción.

histonas presentes en el core, lo que permite filtrar el perfil de RMSD de manera de no tener en cuenta las regiones desestructuradas correspondientes a todos los extremos N-terminales que tendrán mayor variabilidad conformacional esperada. Al hacer estos cálculos se obtiene un valor promedio a lo largo del microsegundo de producción de 0.3 nm donde el valor mínimo observado corresponde a 0.39 nm y el máximo a 0.46 nm, tal como se puede apreciar en la figura 11D.

Dado el tamaño del sistema, y sabiendo que la mayor parte de las interacciones se dan a través de algunos aminoácidos en particular, resulta interesante seguir la evolución de los mismos a lo largo de la dinámica. Para ello se decidió cuantificar aquellos aminoácidos mayoritarios que se encuentran formando parte de la interfase ADN-proteína en función del tiempo. Estos resultados se muestran en la figura 13 donde se representan el número de aminoácidos de arginina, lisina, serina y treonina presentes en la interfase normalizados respecto al valor inicial en la interfase cristalográfica. En la leyenda figura entre paréntesis los valores iniciales de contactos de dichos aminoácidos.

Algunas interacciones caracterizadas por Davey et al. reportan que las regiones predominantemente básicas hacia el extremo amino terminal de las histonas H2B facilitan la estabilización de la estructura del nucleosoma por el sitio de salida al estar interactuando en el surco menor de cadenas adyacentes de ADN. En la figura 14 se muestran los aminoácidos del extremo amino-terminal de H2B que están en los surcos menores del ADN. Se puede apreciar la alta cantidad de residuos básicos (coloreados en azul) presentes en esa zona, y en particular los que están en contacto con el surco menor. También pueden apreciarse 2 lisinas que establecen contactos cada una con el *backbone* de las hebras de ADN adyacentes a la región.

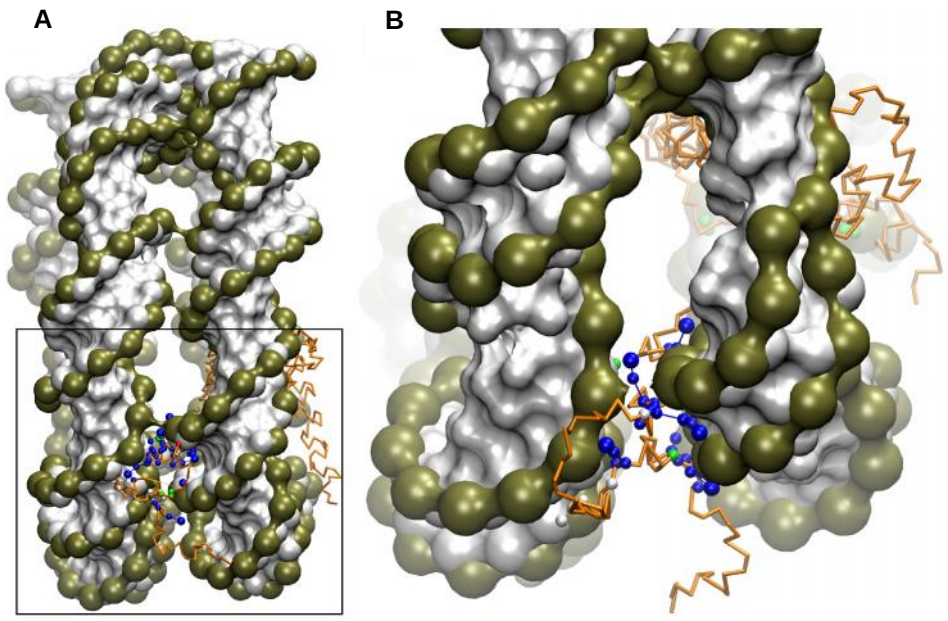


Figura 14: Core del nucleosoma visto de lado con H2B representada por su *backbone* en naranja. La superficie accesible al solvente en los fosfatos se representa en ocre para facilitar la identificación de los surcos. En B se observa con más detalle la zona de interacción del extremo N-terminal entre el ADN proveniente de un giro. Los aminoácidos están coloreados por tipo de residuos, lo que facilita la visualización de los residuos básicos (azul).

Finalmente, al calcular el

porcentaje de conservación de los contactos globales se obtuvo una conservación de 23 % (s.d. 3) con accuracy de 20% (s.d. 3) mientras que los contactos locales alcanzaron valores de 33% (s.d. 5) de conservación y 18 % (s.d. 3) de accuracy asociado. Si bien estos valores pueden parecer algo pequeños, un análisis visual permite apreciar que no existen cambios significativos en la estructura. Además debido al protocolo más detallado utilizado para la equilibración si calculamos la conservación de los contactos contra la estructura inicial de la dinámica se obtiene un porcentaje de 58% (s.d. 6) con accuracy de 51%(s.d. 8) y conservación de interacciones secuencia dependientes de 64 % (s.d. 6) con accuracy asociada de 60% (s.d. 8).

4.5.2 Proteína de unión a TATA-box (TBP)

El siguiente ejemplo se eligió por ser una proteína paradigmática que representa un modo de unión diferente a los vistos en el caso anterior del nucleosoma. TBP es un factor de transcripción que reconoce la secuencia TATA en dirección upstream al inicio de la transcripción. Esta proteína se une al ADN a través de motivos de hoja beta en el surco menor produciendo un marcado *bending* en el ADN. También establece contactos directos entre aminoácidos y el backbone de fosfatos del ADN de manera directa o indirectamente a través de interacciones mediadas por agua. En la figura 15 se presentan la interfase del complejo junto con el RMSD en función del tiempo. Se puede observar que desde los 400ns la interfase se mantiene en un valor aproximado de 28 nm^2. Sin embargo, el perfil de RMSD de los fosfatos permanece estable

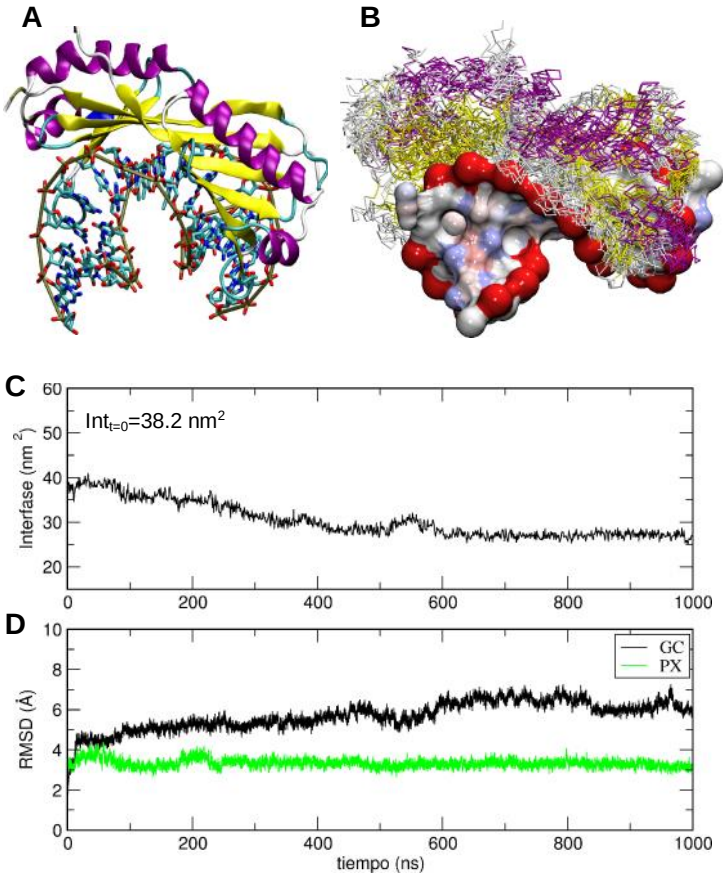


Figura 15: Representación de las estructuras de TBP (PDB ID 1CDW) reportadas en el PDB (A) y a lo largo de la simulación CG (B). En todos los casos la proteína está coloreada por estructura secundaria: alfa hélice (violeta), hoja beta (amarillo), coil (blanco). En B, el ADN está representado con su superficie accesible al solvente aproximada y coloreado por carga: negativo (rojo)-positivo (azul). También se representan varias estructuras simultáneas de la proteína tomadas de la dinámica cada 100ns. En C y D se muestra la superficie de la Interfase ADN-proteína en función del tiempo y en D el RMSD calculado sobre los *beads* GC con estructura secundaria asignada (negro) y sobre todos los *beads* PX del ADN (verde).

desde los 200 ns en torno a 3,5 Å (PX, línea verde) mientras que la proteína experimenta mayores cambios conformacionales oscilando finalmente entorno a valores de 6,3Å. Esto también se ve reflejado en la conservación de contactos. Los contactos globales se conservaron 12% (s.d. 6) con accuracy de 19 (s.d. 4), mientras que los locales, derivados de interacciones con las cadenas laterales reportan valores de conservación de 18% (s.d. 6) y

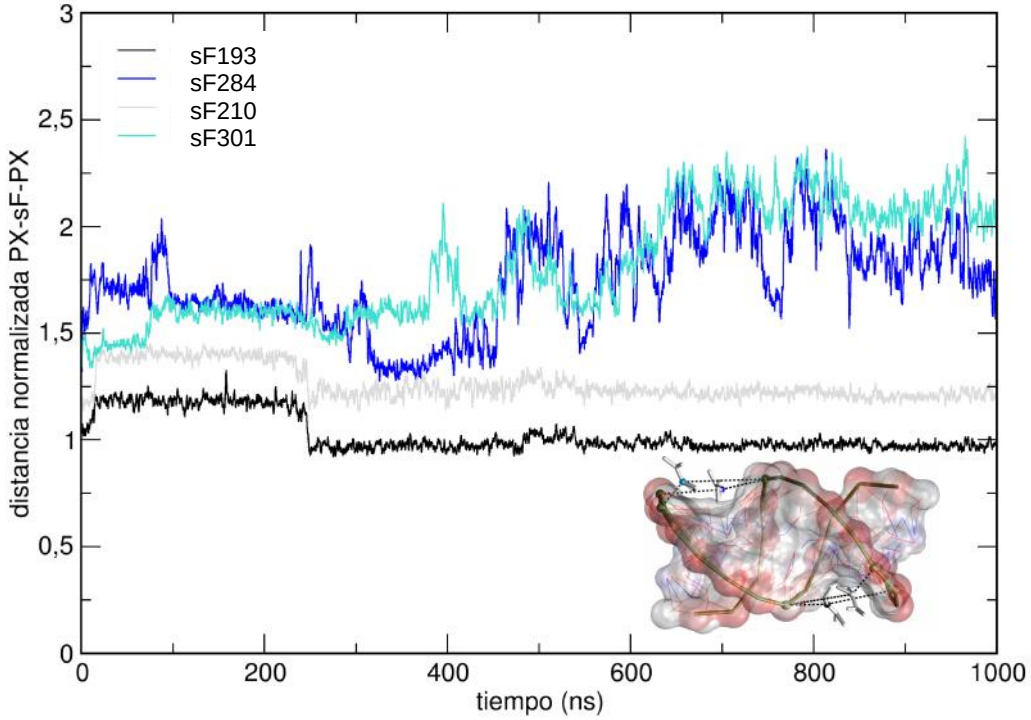


Figura 16: sF193 y sF284 corresponden a las fenilalaninas más insertadas en el surco menor, mientras que sF210 y sF301 corresponden a las fenilalaninas que aportan a la interacción pero se encuentran menos insertadas en el surco menor que las otras dos. SF193 y sF210 (líneas en negro y gris) se encuentran en una zona de interacción mientras que sF284 y sF301 (líneas en azul y turquesa) lo hacen en la región análoga hacia el otro extremo del ADN. El ADN presente en el margen inferior derecho señala los sitios de unión de las fenilalaninas y en negro se marcan las distancias calculadas.

accuracy de 38%. (s.d. 6) Sin embargo, un análisis de cada una de las interacciones de la interfase reportadas en la estructura cristalográfica permite caracterizar con mayor detalle a la interfase.

En la figura 16 se grafican las distancias de interacción representativas de las fenilalaninas insertadas en el surco menor. Para obtener este valor, se sumaron las distancias entre el átomo BCG de la fenilalanina coarse-grained y los PX de las bases correspondientes al par de bases donde se establece el contacto en el surco menor, es decir donde está insertada la fenilalanina. Los contactos pertenecientes a sF193 y sF210 corresponden a la misma zona

de interacción, donde sF193 es el aminoácido más insertado en el surco menor, mientras que sF210 está reportado por cooperar con dicha interacción. Por otro lado, en la región simétrica a este sitio de unión los residuos sF284 y sF301 corresponden también al más insertado en el surco menor y al adyacente. Teniendo en cuenta esto, puede observarse que la zona compuesta por sF193, sF210 y las bases con las que establece contacto (última base de secuencia TATA box, es decir G8, y A7) a pesar de tener oscilaciones representa una menor

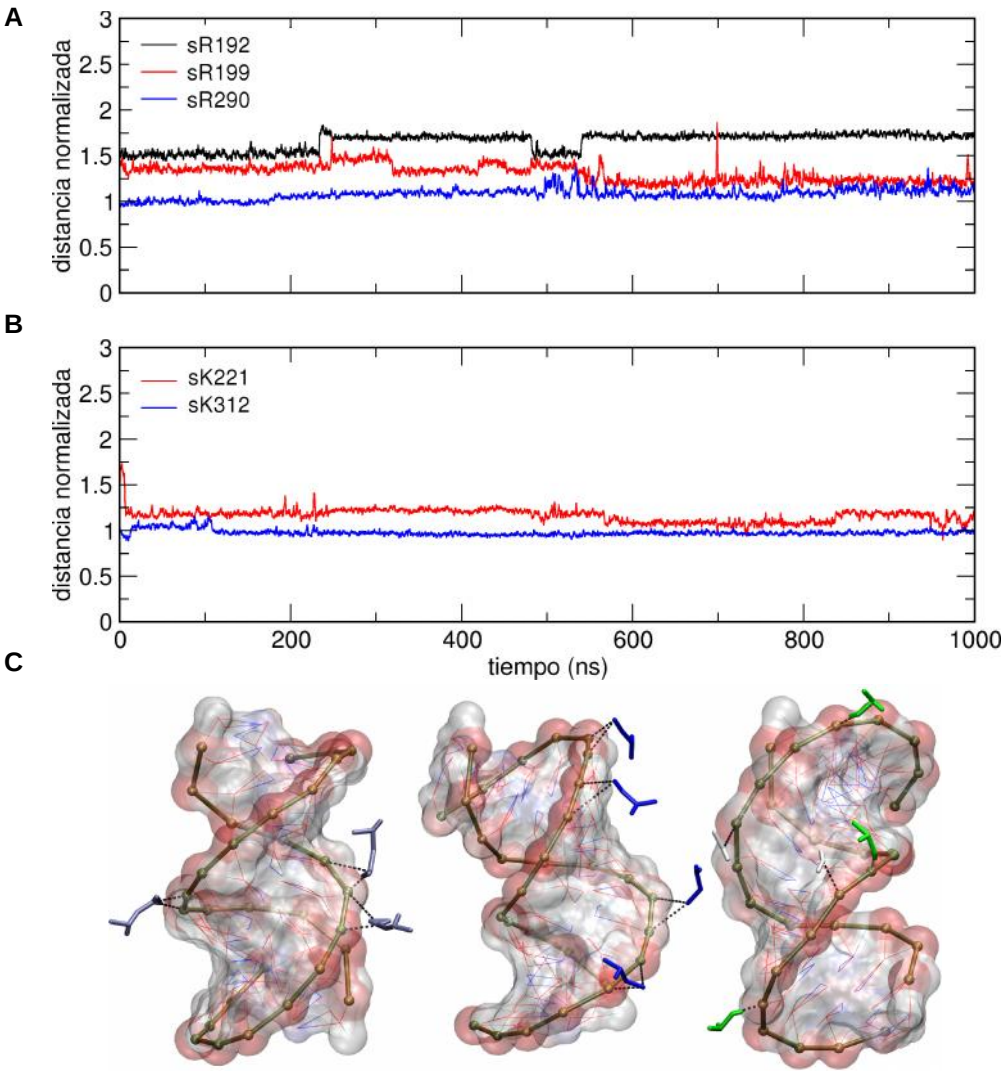


Figura 17: Distancias graficadas de interacciones definidas. A y B muestran valores de distancias normalizados por el valor previo a la minimización de argininas y lisinas respectivamente. En C se señalan las elecciones de distancias a monitorear para el caso de argininas (izq), lisinas (centro) y aminoácidos variados (Ser, Thr, Gly) caracterizados también por interaccionar con el *backbone* de fosfatos del ADN (derecha). En todos los casos el aminoácido de interés está representado como bastones y las distancias señaladas entre los fosfatos son aquellas consideradas para los cálculos.

variabilidad respecto al valor inicial al igual que sF210 en comparación con la otra zona de inserción de fenilalaninas. En esa otra zona el valor inicial de distancia representativa correspondiente a la fenilalanina más enterrada en el surco (sF284) se aleja a valores cercanos a 1.75 veces del inicial mientras que sF310 acaba con un promedio final del entorno del doble del inicial. Estos valores sugieren claramente un modo diferencial de unión de la TBP sobre TATA box desde ambos dominios de la TBP. Otro tipo de interacciones pertenecientes a la interfase de este complejo corresponden a las interacciones existentes a nivel del *backbone* del ADN. En el caso de estas interacciones se optó por representar de manera análoga a lo visto anteriormente, la suma de las distancias entre BCZ (centro de giro más expuesto de la arginina CG, ver tabla 2.2) y ambos fosfatos entre los que se da la interacción. En la figura 17 se muestra la elección de este valor conjuntamente con las gráficas donde se representan únicamente las interacciones directas de lisina o arginina con fosfato. En el caso de la arginina se pueden observar los perfiles estables que poseen estos tres aminoácidos a lo largo de la dinámica, donde en el caso de sR192 parece oscilar entre dos estados pero continua manteniendo valores siempre cercanos a la distancia representativa inicial. Asimismo, para las distancias reportadas de lisinas puede observarse un perfil también estable a lo largo de la dinámica, lo que se relaciona con lo visto anteriormente para argininas.

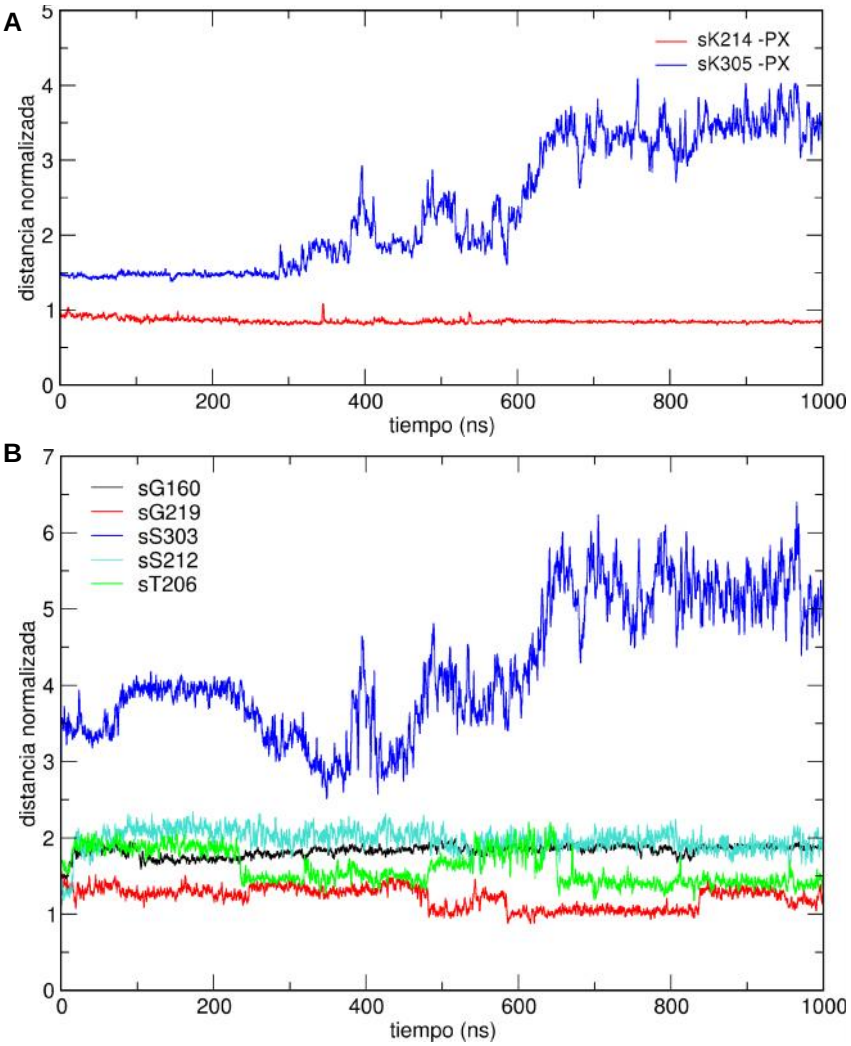


Figura 18: Interacciones reportadas mediadas por agua. En A se muestran aquellas que se dan en el backbone del ADN que incluyen a residuos de lisina. En B se muestran las distancias entre interacciones de residuos polares y glicinas reportadas por los autores en función del tiempo.

Por otro lado, Nikolov et al, también reportaron interacciones mediadas por agua entre

sK214 y sK305 y el *backbone* de ADN. Las distancias totales representativas de estos pares de interacciones se muestran la figura 18, siguiendo el mismo criterio que las interacciones sin mediar agua. En este caso se puede apreciar que sK214 se mantiene estable a lo largo de toda la dinámica mientras que sK305 se desvía notoriamente del valor inicial calculado. Esto va de la mano de lo observado anteriormente para sF284 ya que sK305 se encuentra en la proximidad de sF284 donde la proteína presenta una región con *bends* y tiende a adoptar conformaciones más favorables en esa región. Los autores también reportan interacciones de otros aminoácidos con el *backbone* del ADN. En este caso se trata esencialmente de aminoácidos polares como la serina o treonina así como también alguna interacción establecida por glicinas. En la figura 18 B se presentan los perfiles de distancias entre átomos de las cadenas laterales de dichos aminoácidos y fosfatos. En términos generales se puede comentar que dichas interacciones se mantienen en un estado de aparente equilibro a lo largo del microsegundo a excepción de la serina 303 (sS303). Esta serina al igual que sK305 también está presente en la región C-terminal de la TBP estudiada y corresponde con la región de mayor variación conformacional de la proteína en comparación con el dominio simétrico compuesto por los residuos 155-246, que es el que mantiene también la mayor cantidad de contactos iniciales.

4.5.3 Factor de integración al hospedero (IHF)

Esta proteína es interesante de estudiar dado que reconoce su sitio de unión a través de un mecanismo de lectura indirecta del ADN, lo que implica que reconoce a su blanco principalmente a causa de la estructura que tenga el mismo. IHF es una proteína reportada como capaz de producir *bending* en el ADN, donde las interacciones se dan principalmente a nivel del *backbone* del ADN así como en el surco menor. Cabe destacar que el sistema simulado presenta una curvatura extrema en la secuencia de 35 pares de bases en forma de U así como la presencia de un *nick* en la posición 15, que según reportaron los autores no afecta la estructura global del complejo ni la afinidad del mismo de manera significativa (Lynch, Read et al. 2003). En las gráficas de la figura 19 se puede observar que existe una disminución de la superficie de la interfase ADN-proteína de aproximadamente 5nm^2, manteniéndose estable a partir de los 350 ns de simulación sin restricciones. Esta disminución

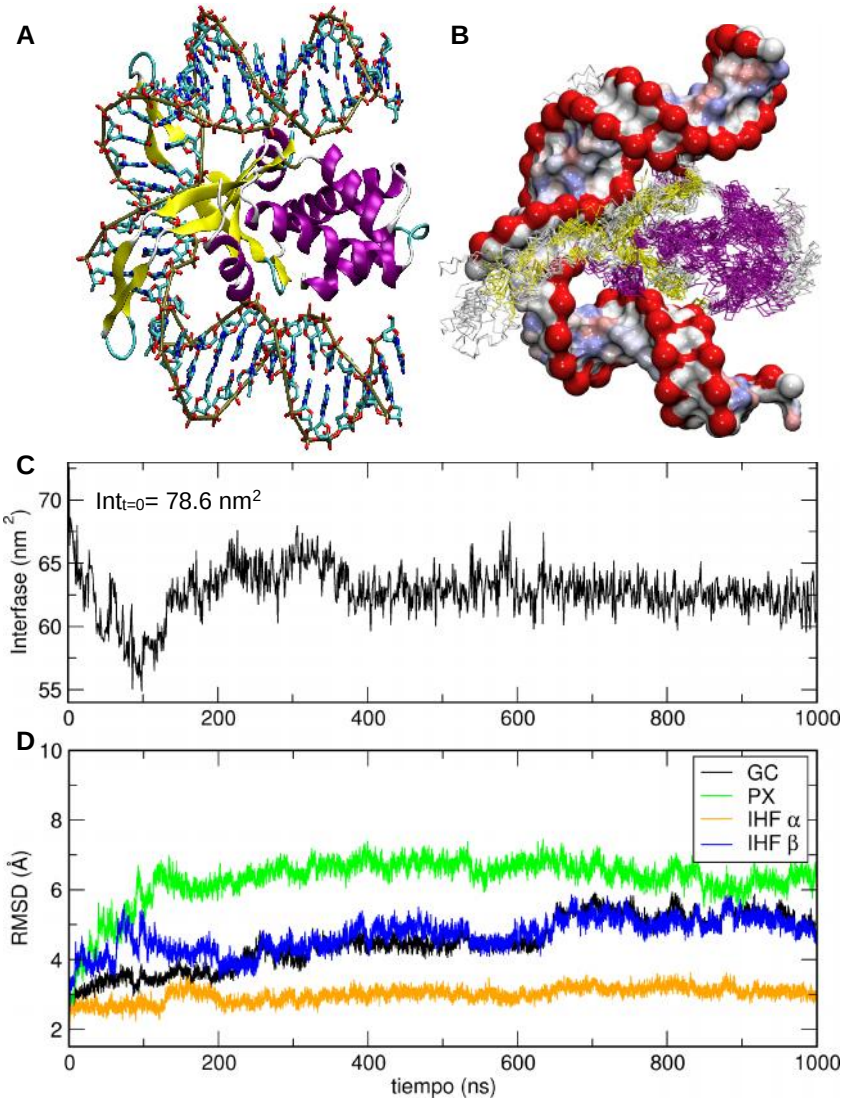


Figura 19: Representación de la estructura del factor de integración al hospedero reportada en el PDB con el código 1OUZ (A) y a lo largo de la simulación CG (B).En ambos casos la proteína está coloreada por estructura secundaria: alfa hélice (violeta), hoja beta (amarillo),*coil* (blanco). En B, el ADN está representado con su superficie accesible al solvente y coloreado por carga: negativo (rojo)-positivo (azul). También se representan varias estructuras simultáneas de la proteína tomadas de la dinámica cada 100ns. En C se muestra la superficie de la Interfase ADN-proteína en función del tiempo y en D el RMSD calculado sobre los *beads* GC con estructura secundaria asignada (negro) y sobre todos los *beads* PX del ADN (verde) y sobre los GC de la subunidad α y β (naranja y azul)

corresponde con la ligera separación de los extremos terminales del ADN al principio de la dinámica, lo que hace perder contactos iniciales mayormente inespecíficos en dichas regiones. El RMSD por lo tanto, presenta un perfil más estable para los átomos de fosfato (PX) a partir de ~200 ns oscilando en valores en torno a 6 Å. Por otra parte si realizamos el análisis de RMSD sobre todos los carbonos alfa de la proteína, sin discriminar las subunidades obtenernos un perfil que acaba estabilizándose en valores cercanos a 5 Å. Sin embargo, si el análisis se realiza sobre los GC de elementos estructurados de cada una de las subunidades por separado se observa que IHFβ presenta un perfil similar al observado para el análisis global donde presenta un crecimiento más rápido en los primeros 100 ns para alcanzar valores finales de 5 Å mientras que IHFα se presenta mucho más estable oscilando en torno a 3 Å. A su vez si analizamos los contactos globales o generados por cadenas laterales para la interfase total entre ADN y proteína obtenemos valores de conservación de 20% (s.d. 4) con accuracy asociada de 35% (s.d. 5) y valores de conservación de 32% (s.d. 4) con accuracy asociada de 41% (s.d. 4) respectivamente. Si en cambio, analizamos los contactos en el caso de IHFα únicamente, obtenemos una conservación de contactos globales de 23% (s.d. 5) con accuracy de 42% (s.d. 8) mientras que para contactos derivados de cadenas laterales estos valores alcanzan 44% (s.d. 5) y 56% (s.d. 5) respectivamente.

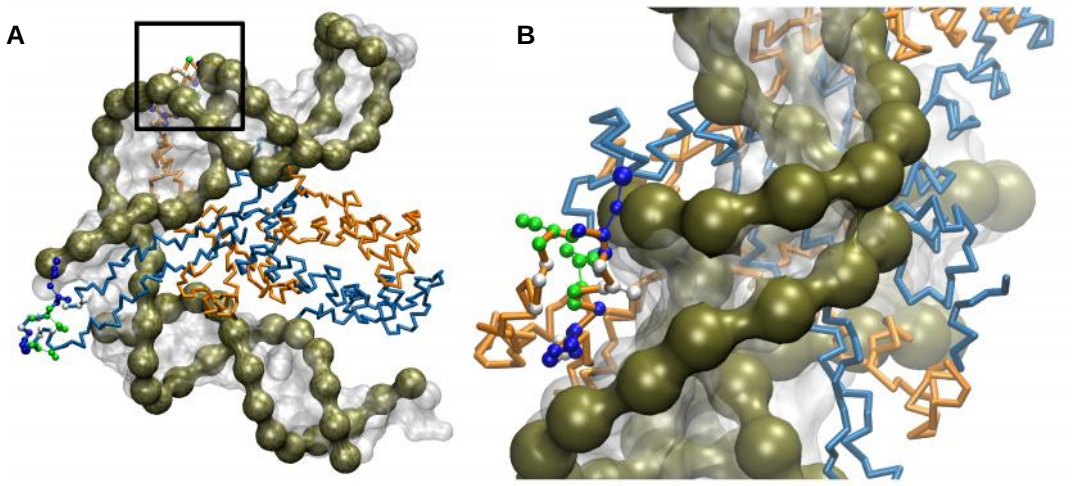


Figura 20: Complejo IHF/ADN. Se representan el *backbone* de la subunidad α de IHF en naranja y el β en azul. El recuadro en A señala la región estudiada en B, donde se puede apreciar los aminoácidos de IHFα que interaccionan en el surco menor y *backbone* del ADN coloreados por tipo de aminoácido (azul:básico, verde:polar, blanco:apolar)

La secuencia aminoacídica GRNPKTG se encuentra conservada en ambas subunidades de IHF presentes en esta estructura e interacciona directamente sobre el surco menor del ADN en la región comprendida por la secuencia ATAGTT tal como se aprecia en la figura 20 donde se muestra a la estructura del complejo proviene de un análisis de cluster realizado sobre todas las estructuras generadas durante la dinámica de 1 μs. Aquí se muestra la estructura más representativa del cluster más poblado y se puede observar que el contacto de esta región con el ADN se mantiene estable en IHFα (región naranja) mientras que se pierde

la intensidad del mismo para IHFβ (región azul). En IHFβ queda prácticamente la lisina en la interfase, mientras que IHFα se puede apreciar que tanto lisina como arginina (azul) se encuentran en el surco menor así como los aminoácidos polares serina y treonina (verde). Este comportamiento explica la diferencia en RMSD observada para cada una de las subunidades y probablemente esté favorecida también por la presencia del *nick* que deja una región con menor densidad de potencial positivo del ADN debido a la ausencia de fosfatos en la región de interfase con IHFβ.

Otra interacción reportada como importante en el reconocimiento indirecto del ADN por esta proteína comprende la región TTR ubicada entre los pares de bases 43 y 45 de la secuencia del bacteriófago λ. En la figura 21 se muestra esta región mapeada en amarillo sobre la superficie accesible al solvente del ADN donde se puede apreciar que la arginina representada en azul se encuentra insertada en el surco menor. La distancia normalizada se presenta estable a lo largo de la dinámica después de los 180 ns aproximadamente en valores en torno a 0.90 lo que expresa una buena conservación de la interacción. La presencia de esta interacción estabilizadora del complejo derivada de la subunidad IHFβ es única pero no basta para optimizar todas las interfases presentes entre esta subunidad y el ADN, lo que genera el aumento en 2 Å del RMSD de esta subunidad en comparación con IHFα.

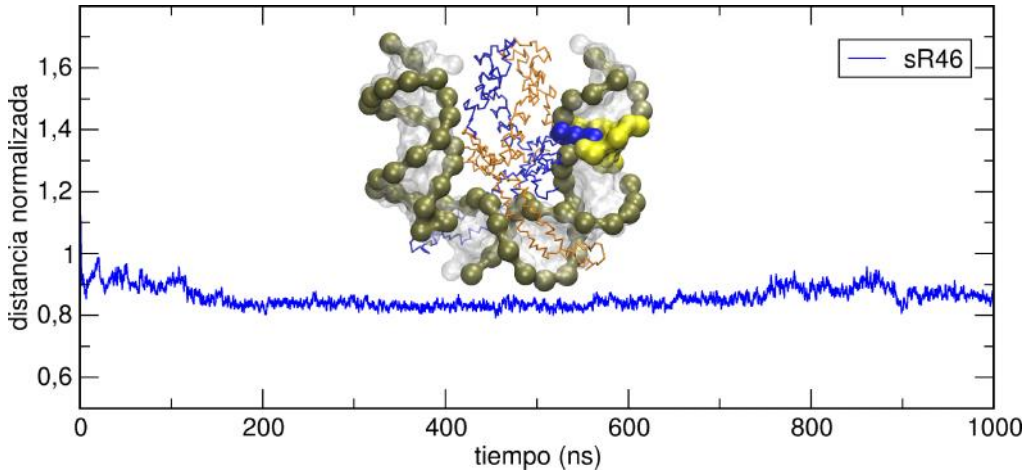


Figura 21: Interfase IHF /TTR.: La interacción de Arg46 de IHFβ (región marcada como esferas de vdW azules) con el ADN se da en la región TTR marcada en amarillo. La distancia entre el *bead* BCG de sR y los fosfatos de ambas hebras de ADN entre los que se ubica en el surco menor fueron sumadas y se siguió su evolución a lo largo de la simulación normalizando respecto al valor inicial derivado de la estructura cristalográfica. Las cadenas IHFα y IHFβ presentan la misma coloración que en la figura anterior.

Finalmente, es posible resumir que, si bien es cierto que se pierden unos pocos puntos de interacción en la subunidad β, el hecho de obtener valores de RMSD de IHFβ en el entorno de 5 Å para una estructura tan compleja está dentro de los valores aceptables para una dinámica *coarse-grained* libre de restricciones.

4.5.4 Resumen final hecho sobre todos los sistemas mostrados

A modo de resumen se presentan a continuación los datos de conservación de contactos globales y de cadenas laterales para todos los sistemas de proteína-ADN simulados excluyendo los leucine zippers.

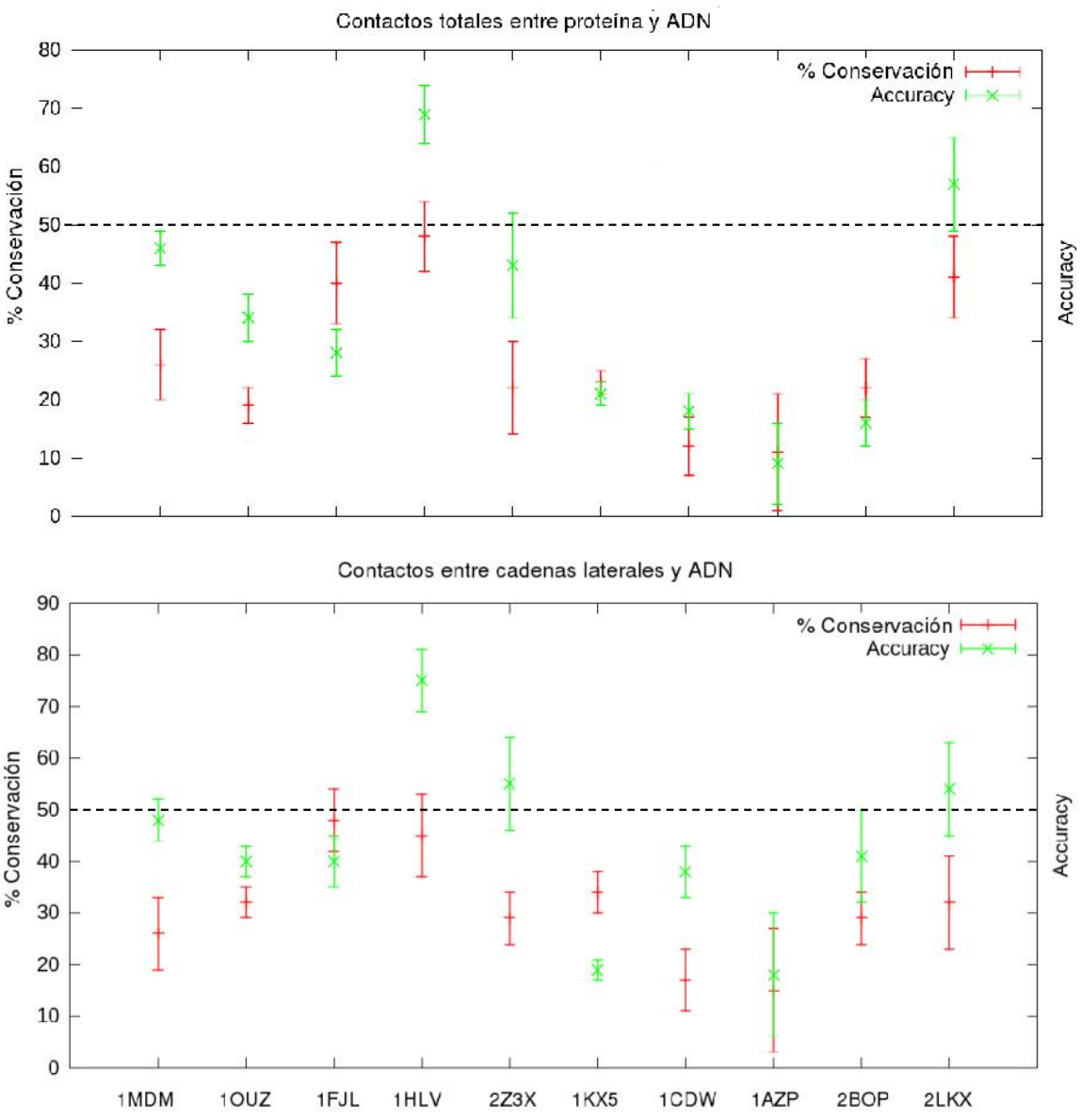


FIgura 22: Conservación y accuracy de contactos para complejos ADN-proteína simulados sin considerar leucine zippers. Arriba se muestran los contactos globales y abajo aquellos que suceden con la cadena lateral únicamente. En ambos casos se representan los valores promedio a lo largo de la dinámica de 1 μs con sus respectivos desvíos estándar.

En el caso de estos sistemas de proteínas variadas es importante remarcar que existen casos donde el ADN está presente con estructuras que distan considerablemente de la forma B canónica, como con los casos de la TBP o Sac7d (PDB ID 1AZP). Dado que el modelo CG de ADN está pensado para representar principalmente B-ADN (permitiendo pequeñas alteraciones locales o globales, tales como *bending* o *narrowing)* en el caso de que éstas se manifiesten marcadamente, es posible perder en cierto grado la estructura inicial y tornarse hacia la forma B. Esto es visible en cierta medida en el caso del complejo TBP/ADN, donde el RMSD global del sistema (considerando todos los átomos GC y PX) presenta picos de hasta 8Å con respecto a la estructura resuelta por cristalografía mientras que el RMSD calculado únicamente sobre PX reporta picos de 4.3Å estabilizándose en valores entorno a 3.5Å. Sin embargo, un análisis detallado del sistema revela que a pesar de la posible pérdida de contactos iniciales presentes en las estructuras cristalográficas, numerosas interacciones características todos los sistemas presentados se mantienen a lo largo de la dinámica.

V DISCUSIÓN

En este trabajo se presentó un primer desarrollo de modelo simplificado para estudiar complejos de ADN-proteínas. Primeramente se compararon las propiedades dinámicas de la estructura de moléculas de ADN doble hélice en presencia de iones monovalentes con datos reportados por cristalografía. Allí se observó que las ocupaciones de los iones a lo largo del surco menor se correspondían con posiciones reportadas en la densidad electrónica, el hecho de haber partido de un modelo generado en la forma de B-ADN canónico desde la secuencia de referencia y haber alcanzado observar eventos de *narrowing* como los presentes en la estructura cristalográfica de referencia demostraron la flexibilidad del modelo y el potencial uso para representar estructuras dinámicas del ADN. Estos resultados permitieron, por lo tanto, continuar con la evaluación de SIRAH para complejos ADN-proteína.

Siguiendo la línea del resultado anterior se planteó la hipótesis, que, puesto que se observó la correcta localización de algunos iones en el surco menor podría también existir localización de argininas o lisinas en las regiones de complejos ADN-proteína con *narrowing* y efectivamente poder mantener el *narrowing* en presencia por ejemplo de argininas. Al hacer las dinámicas de B-ADN en presencia de cadenas laterales de lisina y arginina (figuras 5 y 6) se vio la necesidad de generar interacciones específicas entre átomos de ambos compuestos biológicos. Para ello se decidió realizar un enfoque minimalista corrigiendo estas interacciones. Este enfoque fue necesario para permitir un análisis gradual debido a que al existir 19 atomtypes correspondientes a *beads* de ADN o 39 de proteína totalizan más de 700 interacciones posibles entre ambos tipos de moléculas (sin diferenciar los atomtypes por residuos). Usando las cadenas laterales de lisina y arginina en solución fue posible muestrear con mayor facilidad todas aquellas zonas con interacciones favorecidas principalmente por la electrostática y sin impedimentos estéricos ya que no se consideró el efecto del resto de la proteína en la interacción. A lo largo de la dinámica realizada con las correcciones se observaron las interacciones de arginina y lisina a nivel del *backbone* del ADN (principalmente en regiones interfosfato) con algunas ocurrencias en el surco menor. En particular para la arginina se llegó a observar un estrechamiento del surco menor a causa de la interacción del grupo guanidinio en el mismo. De cualquier manera, al usar cadenas laterales se tuvieron que modificar también los parámetros de los Cα de las cadenas para que no tuvieran interacciones artificiales hidrofóbicas en el surco menor o mayor y así facilitar la correcta orientación de las cadenas laterales respecto al ADN. Además la comparación de las distribuciones de distancias respecto a información representativa del universo de complejos ADN-proteína reportados en el PDB obtuvo perfiles aceptables por lo que se decidió seguidamente cambiar el sistema a estudiar a complejos con proteínas de tipo leucine zipper.

El análisis en la sección 4.4 fue similar al anteriormente descripto, con la salvedad que se introdujeron cambios en las interacciones entre los *beads* que representan al nucleótido en

el surco mayor y los del *backbone* de la proteína y eliminando las correcciones sobre GC de cadenas laterales. Esto se implementó por compatibilidad de modelos de ADN y proteína como se explicó con anterioridad. Se corrieron dinámicas de múltiples sistemas de leucine zippers que fueron elegidos por representar varias interfases de complejos y no ser redundantes siguiendo el protocolo de Norambuena et al(Norambuena and Melo 2010). Los análisis de cluster se realizaron al finalizar la dinámica con un criterio de separación de grupos de RMSD de 3Å ya que en dinámicas CG estables de proteínas el RMSD se estabiliza oscilando entorno a 3Å con respecto a la estructura inicial de referencia. Se guardaron las coordenadas de la estructura más representativa de cada cluster, usando para el análisis tantas estructuras por sistema como fuera necesario para que se representar más del 95% de las conformaciones visitadas durante la dinámica. Este filtro fue importante ya que así se pudo comparar en pie de igualdad a las estructuras resultantes de la dinámica con las cristalográficas originales del mismo set. En la figura 9 se puede observar la mejor compatibilidad de las distribuciones de distancia con las interacciones modificadas al obtener los datos desde proteínas completas y desde el mismo set, tanto para lisina como arginina. Además los datos resultantes del análisis de conservación de contactos arroja valores promedio de 48% de conservación global y 40 % de accuracy, mientras que la conservación local aumenta a 56% con 52% de accuracy, lo cual dada la estricta definición de contacto son resultados favorecedores. Si comparamos estos valores con los del complejo trimérico HP1 Chromo-Shadow/CAF1, la región de reconocimiento entre los monómeros de CS, se conservaba un 55% con accuracy de 26%, lo que comparado con los resultados de los leucine zipper con ADN sugiere, que al menos este último tipo de contacto se comporta de manera más específica dentro de la flexibilidad de la interfase, dando resultados favorables para este tipo de interfase ADN-proteína.

Consecuentemente con la línea de desarrollo planteada se procedió a estudiar sistemas de complejos con mayores desafíos estructurales como se mostró en detalle en la sección 4.5. Considerando los resultados obtenidos en el análisis general de los 10 sistemas simulados es pertinente aclarar que existen algunos casos donde interacciones específicas mediadas por agua se estén perdiendo dada las características del modelo. Desafortunadamente, este es un motivo muy común de interacciones en el surco mayor como sería el caso de la proteína con dominio homoebox Pax (PDB ID 1FJL), donde la hélice que reconoce el ADN media los contactos con el surco mayor, tal como está reportado en la estructura cristalográfica. Por otra parte, en el caso de la proteína cromosomal hipertermófila (Sac7d) la existencia del *kink* no pudo ser mantenida a lo largo de la dinámica. Un modo común para estabilizar las regiones con *kinks* es a través de interacciones hidrofóbicas entre proteínas y las bases que hayan quedado desapareadas a causa de la modificación estructural. En este caso, las interacciones hidrofóbicas no han sido corregidas y podría explicar que en este caso no se pudiera estabilizar lo suficiente al ADN, particularmente por lo corto que es. Sin embargo, dentro de los resultados presentados existe también un caso donde el sistema formado por el factor de integración al hospedero-IHF a pesar de presentar un corte en la molécula de ADN y

adoptar una configuración extrema de forma de U puede mantenerse con cierto grado de estabilidad a lo largo de la dinámica a través de interacción en el surco menor del mismo tipo que las presentas en la estructura cristalográfica. La viabilidad de esta simulación de este complejo que presenta un *nick* se debe probablemente a que el quiebre en el contexto de esta larga molécula logra estabilizarse y genera menor tensión en el ADN total.

En todos estos sistemas el promedio de conservación fue de 27% de conservación global con 37% de accuracy, mientras que la conservación local aumenta a 31% con 43% de accuracy, lo cual es una marcada diferencia con los datos obtenidos únicamente mediante el subset de leucine zipper. De cualquier manera, es esperable este comportamiento ya que las dinámicas en 4.3 son marcadamente más estables que las dinámicas estudiadas en esta sección debido a los *folding* de las diferentes proteínas. Resulta interesante entonces obtener un valor de referencia que dé cuenta de la variabilidad intrínseca de estructuras experimentalmente resueltas. Para ello se realizó un cálculo sobre un complejo ADN-proteína de tipo homeodominio resuelto por NMR (PDB ID 2LKX) y la conservación de contactos entre los confórmeros, utilizando como referencia la estructura más representativa designada por los autores reportó valores de conservación global de 79% con accuracy 77% y conservación local 61% con accuracy de 49%. Esa misma proteína fue simulada por 1 μs obteniendo valores de conservación total de 42% y local de 55% con accuracies asociadas de 58% y 55% en el caso de una proteína con motivos HTH que resulta difícil de estabilizar mediante el modelo CG. Hubiera sido particularmente interesante obtener estos valores de variabilidad para estructuras complejos ADN con leucine zippers y así comparar directamente contra los resultados del subset, sin embargo, hasta el momento no existen estructuras de leucine zippers en complejo con ADN resueltas por NMR en el PDB. De cualquier manera, el hecho de tener una noción experimental de la conservación de contactos puede ayudar a la interpretación de valores obtenidos durante una dinámica molecular. Además es preciso tener en cuenta que si bien los análisis tienen una base en conservación de propiedades estructurales, numerosos mecanismos de reconocimiento se dan desde un punto de vista dinámico. Por lo tanto regiones flexibles en las proteínas como lo son los dominios globulares POU que se encuentran separados por un linker son candidatas a presentar una mayor variabilidad estructural al interactuar con el ADN puesto que la interacción de la región desestructurada en el surco menor regula la separación de ambos dominios estructurados (Fuxreiter, Simon et al. 2011).

VI CONCLUSIONES Y PERSPECTIVAS

Como primer objetivo del trabajo se participó en la optimización y verificación del campo de fuerza actual CG de proteínas (v. ANEXO II). En particular se implementaron correcciones a nivel de interacciones entre pseudoátomos capaces de formar puentes salinos, así como también se elaboraron estrategias de análisis para las dinámicas de proteínas.

Además, el estudio de la dinámica molecular *coarse-grained* de B-ADN en solución salina (sección 4.1) arrojó evidencias de la capacidad del campo de fuerzas de reproducir estructuras cristalográficas de manera dinámica. En particular, se observó que los cationes se ubican a lo largo del surco menor ocupando posiciones comparables con densidades electrónicas reportadas.

Esta evidencia facilitó la idea de probar la compatibilidad del campo de fuerza de proteínas recientemente desarrollado con la versión ya existente de ADN. Puesto que ambos fueron desarrollados con el mismo criterio de representar estructuras experimentales todos los análisis se realizaron en base a estructuras resueltas experimentalmente. La comparación de distribución de distancias de arginina y lisina contra ADN hecha con datos del PDIdb y simulaciones iniciales reveló la necesidad de corregir dichas interacciones. Esto fue necesario principalmente a causa de la elevada interacción electrostática proveniente de las altas cargas localizadas en los átomos CG. Una vez implementada la corrección en sistemas compuestos por ADN y cadenas laterales de dichos aminoácidos se decidió probar dichos parámetros en sistemas con cadenas polipeptídicas de proteínas estructuradas como los leucine zippers.

Las simulaciones realizadas en complejos compuestos por leucine zippers arrojaron valores aceptables de conservación de contactos y accuracy, aunque demostraron una ligera tendencia a generar contactos *de novo* no existentes en las estructuras iniciales de referencia. Debido a que los sistemas con leucine zippers no pueden brindar información representativa del universo de posibles modos de interacción ADN-proteína (a nivel del surco menor, por ejemplo), se simularon una serie de complejos que abarcaran mayor tipo de interacciones. Estos resultados mostraron que, efectivamente la complejidad del conjunto elegido es notoriamente mayor ya que, en términos generales se obtuvieron resultados menos favorecedores. De cualquier manera, estructuras donde el ADN se presente mayormente en forma B tienden a ser más estables y en particular aquellas donde existen interacciones puntuales de argininas en el surco menor también reproducen resultados satisfactorios, incluso manteniendo el *narrowing*. Por otro lado, sistemas extremos con el Sac7 donde una serie de hélices alfa perteneciendo a diversos monómeros interactúa en el surco menor genera distorsiones considerables en el ADN y en el complejo también. Asimismo, estructuras con motivos HTH donde en los *turns* existan conformaciones de enlaces peptídicos en *cis* pueden generar distorsiones importantes en el *folding* de la proteína ya que en la versión actual del

campo de fuerza no existen parámetros que los representen.

A su vez, es importante considerar que las modificaciones implementadas en el presente trabajo mejoran notoriamente la estructura de los complejos estudiados, de cualquier manera interacciones como intercalación de aminoácidos hidrofóbicas entre las bases nitrogenadas del ADN no pueden ser reproducidas en la presente versión aunque sí pueden conservarse interacciones hidrofóbicas a lo largo del surco menor (esto se vio con fenilalaninas en el ejemplo TBP/ADN en la interfase conservada del semieje N-terminal de TBP, figura 16) Por supuesto, es importante que el usuario sea consciente de las limitaciones intrínsecas del modelo utilizado como en casos de proteínas con *turns* o ADN en forma Z o con distorsiones extremas para los que el modelo no fue diseñado.

Finalmente se puede concluir que con las interacciones corregidas es posible simular complejos ADN-proteína adecuadamente y en particular si tienen motivos de interacción que no tengan elementos limitantes del campo de fuerza. De cualquier manera, existe el interés de seguir trabajando para mejorar la representatividad de las interfases, teniendo ya proyectos programados para expandir el muestreo y complementariedad con la siguiente versión del campo de fuerza que está en desarrollo.

VII AGRADECIMIENTOS

Quisiera agradecer a la Agencia Nacional de Investigación e Innovación (ANII) por haberme financiado la beca de Maestría en este tiempo y a PEDECIBA Bioinformática por el apoyo, en particular la beca para asistencia a congresos en el exterior.

Especialmente quiero agradecer a mi orientador, Sergio Pantano por todo el apoyo y la guía en este trabajo así como en mis comienzos como estudiante de grado. También me gustaría agradecerle a mi co-orientador Francisco Melo por sus sugerencias en el análisis de datos.

A todos los compañeros del Grupo de Simulaciones Biomoleculares con los que compartimos este tiempo de maestría: Leonardo Darré, Matías Machado, Humberto González, Sebastián Ferreira, Gastón Hugo y Steffano Silva.

También agradezco a Andreas Schüller por asistirme con la recopilación de datos de la PDIdb.

A mi familia y los buenos amigos de toda la vida por acompañarme en este camino a pesar de la distancia. Especialmente quiero agradecer a mis padres por el apoyo incondicional desde siempre.

VIII REFERENCIAS

Baaden, M. and S. J. Marrink (2013). "Coarse-grain modelling of protein-protein interactions." Current Opinion in Structural Biology **23**(6): 878-886.

Berendsen, H. J. C., J. P. M. Postma, et al. (1984). "Molecular dynamics with coupling to an external bath." The Journal of Chemical Physics **81**(8): 3684-3690.

Biswas, M., J. Langowski, et al. (2013). "Atomistic simulations of nucleosomes." Wiley Interdisciplinary Reviews: Computational Molecular Science **3**(4): 378-392.

Bussi, G., D. Donadio, et al. (2007). "Canonical sampling through velocity rescaling." The Journal of Chemical Physics **126**(1): 014101.

Chaney, B. A., K. Clark-Baldwin, et al. (2005). "Solution Structure of the K50 Class Homeodomain PITX2 Bound to DNA and Implications for Mutations That Cause Rieger Syndrome." Biochemistry **44**(20): 7497-7511.

Dans, P. D., L. Darré, et al. (2013). Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level. Advances in Bioinformatics and Computational Biology. J. Setubal and N. Almeida, Springer International Publishing. **8213:** 71-81.

Darré, L., M. R. Machado, et al. (2015). "SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics." Journal of Chemical Theory and Computation **11**(2): 723-739.

Darré, L., M. R. Machado, et al. (2010). "Another Coarse Grain Model for Aqueous Solvation: WAT FOUR?" Journal of Chemical Theory and Computation **6**(12): 3793-3807.

Davey, C. A., D. F. Sargent, et al. (2002). "Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 A Resolution" Journal of Molecular Biology **319**(5): 1097-1113.

Dolinsky, T. J., J. E. Nielsen, et al. (2004). "PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations." Nucleic Acids Research **32**(suppl 2): W665-W667.

Fuxreiter, M., I. Simon, et al. (2011). "Dynamic protein-DNA recognition: beyond what can be seen." Trends in Biochemical Sciences **36**(8): 415-423.

Garvie, C. W., M. A. Pufall, et al. (2002). "Structural Analysis of the Autoinhibition of Ets-1 and Its Role in Protein Partnerships." Journal of Biological Chemistry **277**(47): 45529-45536.

Hegde, R. S., S. R. Grossman, et al. (1992). "Crystal structure at 1.7 A of the bovine

papillomavirus-1 E2 DMA-binding domain bound to its DNA target." Nature **359**(6395): 505-512.

Humphrey, W., A. Dalke, et al. (1996). "VMD - Visual Molecular Dynamics " J. Molec. Graphics **14** 33-38.

Karplus, M. and J. Kuriyan (2005). "Molecular dynamics and protein function." Proceedings of the National Academy of Sciences of the United States of America **102**(19): 6679-6685.

Keller, W., P. König, et al. (1995). "Crystal Structure of a bZIP/DNA Complex at 2.2 A: Determinants of DNA Specific Recognition." Journal of Molecular Biology **254**(4): 657-667.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy." Science **302**(5649): 1364-1368.

Leach, A. R., Ed. (2001). Molecular modelling principles and applications Pearson Education Limited.

Lo Conte, L., B. Ailey, et al. (2000). "SCOP: a Structural Classification of Proteins database." Nucleic Acids Research **28**(1): 257-259.

Lynch, T. W., E. K. Read, et al. (2003). "Integration Host Factor: Putting a Twist on Protein-DNA Recognition." Journal of Molecular Biology **330**(3): 493-502.

Lynch, T. W., E. K. Read, et al. (2003). "Integration Host Factor: Putting a Twist on Proteinâ€“DNA Recognition." Journal of Molecular Biology **330**(3): 493-502.

Marrink, S. J., H. J. Risselada, et al. (2007). "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations." The Journal of Physical Chemistry B **111**(27): 7812-7824.

Monticelli, L., S. K. Kandasamy, et al. (2008). "The MARTINI Coarse-Grained Force Field: Extension to Proteins." Journal of Chemical Theory and Computation **4**(5): 819-834.

Nikolov, D. B., H. Chen, et al. (1996). "Crystal structure of a human TATA box-binding protein/TATA element complex." Proceedings of the National Academy of Sciences of the United States of America **93**(10): 4862-4867.

Noid, W. G. (2013). "Perspective: Coarse-grained models for biomolecular systems." The Journal of Chemical Physics **139**(9): 090901.

Norambuena, T. and F. Melo (2010). "The Protein-DNA Interface database." BMC Bioinformatics **11**: 262-262.

Parrinello, M. and A. Rahman (1981). "Polymorphic transitions in single crystals: A new molecular dynamics method." Journal of Applied Physics **52**(12): 7182-7190.

Pasi, M., R. Lavery, et al. (2013). "PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties." Journal of Chemical Theory and Computation **9**(1): 785-793.

Perez, A. and M. Orozco (2010). "Real-Time Atomistic Description of DNA Unfolding." Angewandte Chemie International Edition **49**(28): 4805-4808.

Periole, X., M. Cavalli, et al. (2009). "Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition." Journal of Chemical Theory and Computation **5**(9): 2531-2543.

Pronk, S., S. Páll, et al. (2013). "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit." Bioinformatics **29**(7): 845-854.

Robinson, H., Y.-G. Gao, et al. (1998). "The hyperthermophile chromosomal protein Sac7d sharply kinks DNA." Nature **392**(6672): 202-205.

Rohs, R., S. M. West, et al. (2009)[1]. "Nuance in the Double-Helix and its Role in Protein-DNA Recognition." Current Opinion in Structural Biology **19**(2): 171-177.

Rohs, R., S. M. West, et al. (2009)[2]. "The role of DNA shape in protein-DNA recognition." Nature **461**(7268): 1248-1253.

Schlick, T. (2010). Molecular Modeling and Simulation, An Interdisciplinary Guide.

Shirai, T., Y. Watanabe, et al. (2009). "Structure of Rhamnose-binding Lectin CSL3: Unique Pseudo-tetrameric Architecture of a Pattern Recognition Protein." Journal of Molecular Biology **391**(2): 390-403.

Shui, X., L. McFail-Isom, et al. (1998). "The B-DNA dodecamer at high resolution reveals a spine of water on sodium." Biochemistry **37**(23): 8341-55.

Sterpone, F., S. Melchionna, et al. (2014). "The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems." Chemical Society reviews **43**(13): 4871-4893.

Tanaka, Y., O. Nureki, et al. (2001). Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA.

Wilson, D. S., B. Guenther, et al. "High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA." Cell **82**(5): 709-719.

Zhao, G., J. R. Perilla, et al. (2013). "Mature HIV-1 capsid structure by cryo-electron microscopy

and all-atom molecular dynamics." Nature **497**(7451): 643-6.

ANEXO I Tablas con todos los complejos simulados

Tabla 1 ANEXO I: Resultados de Conservación de contactos y accuracy con sus respectivos desvíos estándar para todos los sistemas presentados tomando como referencia los contactos nativos presentes en la estructura cristalográfica. Se listan en función del código PDB. Aquellos valores con fondo celeste corresponden a los complejos utilizados en la sección 4.5 mientras que los restantes son complejos con proteínas de tipo leucine zipper (sección 4.4)

PDB	Contactos globales		Contactos locales (cadena lateral)	
	Conservación	Accuracy	Conservación	Accuracy
1AM9_p1	43 s.d. 3	34 s.d. 5	49 s.d. 4	50 s.d. 6
1AM9_p2	33 s.d. 6	45 s.d. 7	36 s.d. 6	50 s.d. 8
1GD2	38 s.d. 4	55 s.d. 6	49 s.d. 5	65 s.d. 5
1GTW	42 s.d. 4	36 s.d. 3	65 s.d. 4	51 s.d. 3
1GU5	46 s.d. 6	36 s.d. 5	60 s.d. 7	60 s.d. 5
1GU4	55 s.d. 3	39 s.d. 2	70 s.d. 3	51 s.d. 2
1H8A	61 s.d. 4	42 s.d. 3	70 s.d. 4	55 s.d. 4
1H89	49 s.d. 4	41 s.d. 4	66 s.d. 8	60 s.d. 6
1NLW	44 s.d. 8	32 s.d. 6	49 s.d. 5	50 s.d. 6
1NKP	52 s.d. 4	48 s.d. 6	50 s.d. 7	54 s.d. 9
2C9L	58 s.d. 6	32 s.d. 3	61 s.d. 8	39 s.d. 7
2QL2	63 s.d. 6	28 s.d. 5	60 s.d. 7	59 s.d. 6
2H7H	39 s.d. 4	36 s.d. 5	46 s.d. 5	39 s.d. 6
1NJM	49 s.d. 6	26 s.d. 3	59 s.d. 6	40 s.d. 4
2DGC	46 s.d. 6	53 s.d. 6	54 s.d. 5	50 s.d. 6
2E42	35 s.d. 8	35 s.d. 4	47 s.d. 9	37 s.d. 4
2E43	37 s.d. 5	28 s.d. 3	54 s.d. 6	47 s.d. 3
1KX5	23 s.d. 3	20 s.d. 3	33 s.d. 5	18 s.d. 3
1MDM	27 s.d. 6	47 s.d. 4	26 s.d. 8	48 s.d. 5
1OUZ	20 s.d. 3	35 s.d. 4	32 s.d. 4	41 s.d. 4
1FJL	42 s.d. 8	28 s.d. 5	50 s.d. 7	41 s.d. 5
1HLV	49 s.d. 6	70 s.d. 5	46 s.d. 8	76 s.d. 6
2Z3X	22 s.d. 8	44 s.d. 9	29 s.d. 6	55 s.d. 9
1KX5	23 s.d. 3	21 s.d. 3	35 s.d. 5	19 s.d. 3
1CDW	12 s.d. 6	19 s.d. 4	18 s.d. 6	38 s.d. 6
1AZP	12 s.d. 11	10 s.d. 8	16 s.d. 12	18 s.d. 12
2BOP	22 s.d. 5	17 s.d. 5	30 s.d. 6	42 s.d. 10
2LKX	42 s.d. 8	58 s.d. 8	33 s.d. 10	55 s.d. 9

ANEXO II: Artículos completos publicados a la fecha de presentación de la tesis

Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level

Pablo D. Dans[1,2], Leonardo Darré[1,3], Matías R. Machado[1], Ari Zeida[1,4], Astrid F. Brandner[1], and Sergio Pantano[1,*]

[1] Institut Pasteur de Montevideo, Mataojo 2020, 11400, Uruguay
[2] Institute for Research in Biomedicine (IRB Barcelona), Baldiri Reixac, 10, 08028 Barcelona, Spain
[3] Department of Chemistry, King's College London, London, United Kingdom
[4] Deptartamento de Qca. Inorgánica, Analítica y Química-Física and INQUIMAE-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. 2,C1428EHA Buenos Aires, Argentina
spantano@pasteur.edu.uy
www.sirahff.com

Abstract. We present a comparison between atomistic and coarse grain models for DNA developed in our group, which we introduce here with the name SIRAH. Molecular dynamics of DNA fragments performed using implicit and explicit solvation approaches show good agreement in structural and dynamical features with published state of the art atomistic simulations of double stranded DNA (using Amber and Charmm force fields). The study of the multi-microsecond timescale results in counterion condensation on DNA, in coincidence with high-resolution X-ray crystals. This result indicates that our model for solvation is able to correctly reproduce ionic strength effects, which are very difficult to capture by CG schemes.

Keywords: Molecular dynamics, nucleic acids, simulations, WT4, flexibility, counterions, narrowing.

1 Introduction

Molecular Dynamics (MD) simulations have become a trustworthy and useful tool for the study of the structural and dynamical behavior and interactions between biomolecules [1]. However, despite continuous developments [2-4], this technique is limited to relatively small systems or short simulation times. Therefore, effort has been devoted to the implementation of simulation techniques based on the idea of simplified or coarse grained (CG) representations of atomistic or fine grain (FG) systems, which reduce significantly the computational demands but still capture the physical essence of the phenomena under examination (see ref. [5] for a comprehensive review). Owing to its biological relevance, DNA has been subject of development of several CG

* Corresponding author.

J.C. Setubal and N.F. Almeida (Eds.): BSB 2013, LNBI 8213, pp. 71–81, 2013.

models (recently reviewed in [6]). Among others our group has developed a model for CG simulations of DNA [7] (Fig. 1a, b), which we present here for the first time with the acronym SIRAH (Southamerican Initiative for a Rapid and Accurate Hamiltonian). The SIRAH model can be used in combination with implicit or explicit solvation schemes using generalized Born model or a CG water model called WatFour [8] (WT4 for shortness, Fig. 1c). Moreover, it can be used for dual resolution simulations, in which FG and CG segments can be intercalated within the same DNA filament [9]. Here we present a critical assessment of the accuracy of our model using the Drew-Dickerson dodecamer (DD) [10] as main benchmark to compare our results with FG simulations and/or experimental data. The backmapping of CG coordinates allows recovering pseudo atomistic information from calculations performed at the CG level. Finally, we show that SIRAH can reproduce ionic strength effects and species-specific ionic binding to DNA, which are in agreement with high resolution X-ray data [11], and lead to a significant bending of the double helix.

2 Methods

2.1 The SIRAH Force Field for CG DNA and Aqueous Solvation

Our CG model for DNA (Fig. 1a, b) uses six effective beads per nucleobase, each placed in correspondence with the positions of real atoms in canonical conformations of FG nucleotides. A comparative view of the topology and excluded sizes of nucleotides, CG water and ions is shown in Fig. 1c. The partial charges on each bead add to a unitary negative charge on each nucleotide and ensure Watson-Crick electrostatic recognition. This charge distribution generates dipole moments, which are well compatible with those of state of the art FG force fields (Fig. 1d, e). A complete description of all parameters can be found in refs. [7,8]. In analogy with transient tetrahedral clusters formed by pure water, our model uses four beads interconnected in a tetrahedral conformation (Fig. 1c) [8]. Since each bead carries an explicit partial charge, WT4 liquid generates its own dielectric permittivity without the need to impose a uniform dielectric. The WT4 model reproduces several common properties of liquid water and simple electrolyte solutions [8].

2.2 MD Simulations and Analysis

The SIRAH model runs straightforwardly in the simulation packages AMBER and GROMACS (input/parameter files and tools to convert and visualize molecules are available from www.sirahff.com). Implicit solvent simulations, using the HCT pairwise generalized Born model [12] are performed with AMBER[13] using a cutoff of 18 Å and a salt concentration of 0.15 M. Temperature is controlled using a Langevin thermostat [14,15] with a friction constant of 50 ps^{-1}. Explicit solvent simulations are performed in the NPT ensemble using GROMACS 4.5[16]. A direct cut off for non-bonded interactions of 12 Å is used while long range electrostatics were evaluated using the PME approach [17]. Temperature and pressure are coupled to Berendsen thermostats and barostats [18] with coupling times of 1.2 ps and 6.0 ps, respectively. All systems are energy minimized and stabilized by raising the temperature from 0ºK to 300ºK in 1 ns.

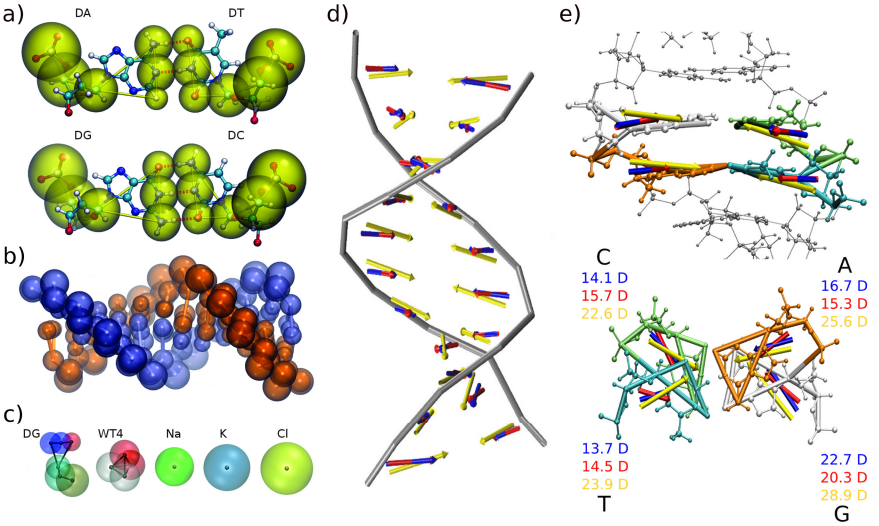


Fig. 1. SIRAH force field for CG DNA and aqueous solvent. a) FG nucleotides are presented with balls and sticks and colored by atoms (Oxygen: red, Nitrogen: blue, Carbon: cyan, Hydrogen: white). CG beads placed on FG positions are shown as semitransparent spheres. b) CG representation of the double helix DNA. c) Comparative view of CG molecules. Guanine, WT4 (water) and CG ions sodium, potassium and chloride are shown from left to right. Single electrolytes implicitly include a first solvation shell. The diameter of the CG beads in (a) and (c) corresponds to the actual van der Waals radii, providing an idea of the effective excluded volume and relative sizes of the beads. d) Schematic representation of dipole moments along the X-ray structure 1BNA showing the colinearity between FG and CG schemes (blue, red and yellow indicate Amber99-bsc0, Charmm27 and SIRAH force fields, respectively). e) Top: dipoles drawn on AT and GC base pairs showing their alignment with the base's planes. Bottom: same as top but seen along the DNA axis. Dipole modules are expressed in Debyes.

A time step of 20 fs is used. A list of the simulated systems and their composition are presented in Tab. 1. The X-ray structure of the Drew-Dickerson dodecamer (PDB id:1BNA [10]) was used as starting point. We compared our results against FG simulations on the same system performed with the Amber99-bsc0 [13] force field reported by Pérez et al [19] (sys3 in Tab. 1, available on-line at http://mmb.pcb.ub.es/microsecond/). Additionally, two systems bearing the 10 unique dinucleotide steps (namely, AA·TT, AC·GT, AG·CT, AT·AT, CA·TG, CC·GG, CG·CG, GA·TC, GC·GC and TA·TA) were simulated starting from the canonical B-form (sys5 and sys6 in Table 1) and compared with results reported in reference [20]. All the comparisons have been made on the back-mapped trajectories according to the procedure described in ref. [7]. Helical parameters are calculated using Curves+[21]. Root Mean Square Deviations (RMSD) are computed on all heavy atoms excluding the capping base pairs, while major and minor groove dimensions are measured between opposite phosphate groups and averaged along the double helix. Eigenvectors and eigenvalues are obtained by diagonalization of the covariance matrix calculated along the trajectories for all the heavy atoms using standard GROMACS utilities. As a gauge of the likeliness between different simulations, trajectories are fitted to

Table 1. Description of the simulated systems

System	Solvation model	nº solvent molecules [a]	Ionic Species (nº of ions)	Nucleotide sequence (5'to 3')	time (μs)
Sys1[c]	GB	---	---	CGCGAATTCGCG [b]	1.2
Sys1wr	GB	---	---	CGCGAATTCGCG [b]	1.2
Sys2[c]	WT4	523 (5753)	Na+(22)	CGCGAATTCGCG [b]	1.2
Sys2wr	WT4	523 (5753)	Na+(22)	CGCGAATTCGCG [b]	12
Sys3[d]	TIP3P	4998	Na+(22)	CGCGAATTCGCG [b]	1.2
Sys4	WT4	506 (5566)	Na+(19) K+(19) Cl-(16)	CGCGAATTCGCG [b]	12
Sys5	WT4	1510 (16610)	Na+(34) K+(33) Cl-(33)	GCCTATAAACGCCTATAA	10
Sys6	WT4	1510 (16610)	Na+(34) K+(33) Cl-(33)	CTAGGTGGATGACTCATT	10

[a] Parenthesis indicate the equivalent number of FG water molecules represented. [b] Drew-Dickerson dodecamer. [c] Simulated using harmonic constraints on the capping base-pairs. [d] Taken from ref.[19].

a common reference (the canonical structure) and their covariance matrices compared using a similarity index (SI) as in ref. [9].

The essential dynamics analysis is performed to compare sys1/sys2 with sys3. To avoid contaminating the main components of motion with possible helix fraying, in sys1 and sys2 constraints of 0.75 Kcal/mol•Å^2 are applied only to the Watson-Crick beads of capping bases. Counterions condensation is analyzed by computing electrolyte occupancy density maps in 3D regular grids of 0.3 Å using VMD [22]. Cations closer than 5 Å to phosphate groups of both opposite strands are considered bound to the minor groove. Narrowing is measured only on the central track, i.e. between the four central phosphate pairs.

3 Results and Discussion

3.1 Structural and Dynamical Comparison

A first comparison of our CG model versus FG simulations (sys3) is performed in terms of RMSD on the backmapped trajectories of simulations using implicit and explicit solvation (sys1 and sys2, respectively). Along the 1,2 μs explored, both solvation schemes described equally well the DD structure with no RMSD drift from the experimental structure (Fig. 2a). In all the simulations the DNA duplexes show a flexible but stable behavior oscillating around the equilibrium B-form. The higher number of conformational substates explored by the FG simulation translates in higher RMSD variations. A good

correspondence between the three simulations can be also inferred from structural superposition of backmapped snapshots taken at the beginning, middle and end of the dynamics (Fig. 2b).

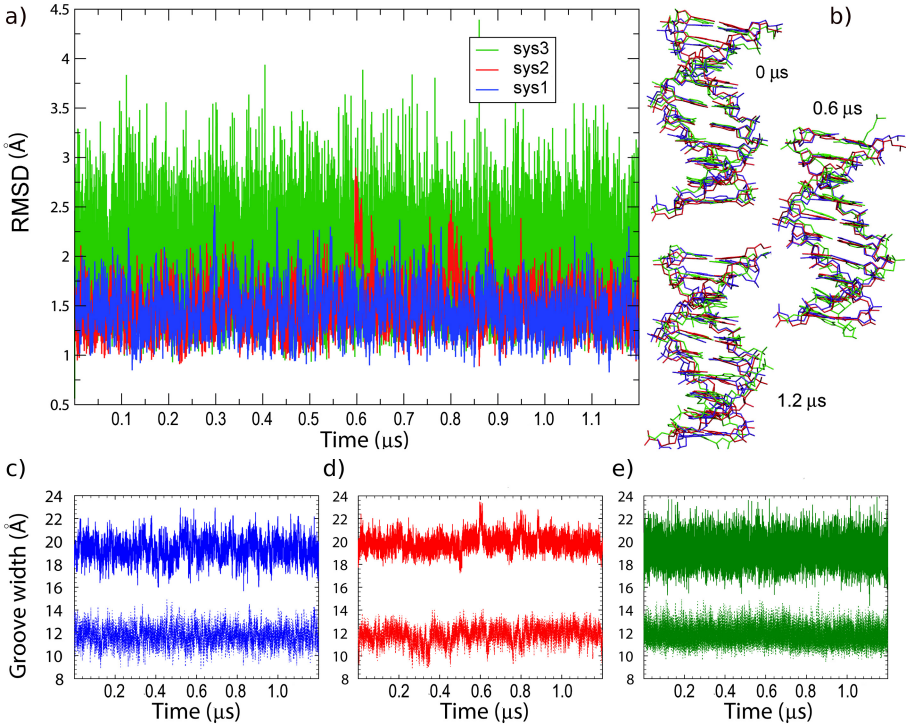


Fig. 2. Comparison between FG and CG simulations of the DD dodecamer. Implicit solvation CG (sys1), explicit solvation CG (sys2) and FG simulations (sys3) are presented in blue, red and green, respectively. a) RMSD along time calculated for all the heavy atoms (FG and backmapped CG) respect to the X-ray structure 1BNA. b) Least mean square fit performed on all heavy atoms of conformers taken at the beginning, middle, and end of the trajectories. c - e) Major and minor groove widths (top and bottom traces, respectively) for the three systems.

Other characteristic features of DNA as the minor and major grooves show also a very good agreement with the FG simulation (Fig. 2c-e). In correspondence with the observation made for RMSD, both CG schemes show lower fluctuations. To gain a deeper insight on the dynamical behavior of the CG simulations, we compare the conformational subspace sampled by inspecting the essential dynamics modes of each simulation. In all the cases, the first 3 eigenvectors explained nearly 50% of the total variance. These 3 essential modes are analyzed in more detail in terms of their projection onto the real space (see animation at http://www.youtube.com/watch?v=ivW7ixsG0fA&feature=youtu.be). The first mode involves a twisting and untwisting, the second is related with a simultaneous bending and twisting around the center of the AT track, while the third eigenvector, is associated to a global tilting of the duplex. To achieve a more quantitative characterization of the likeness between trajectories we calculate a similarity index (SI) from the

covariance matrices of each simulation. As a measure for the maximum similarity reachable during the time window explored, we divided the FG trajectory in two halves and calculate the SI between both segments of trajectory resulting in a value of 0.91. Comparison of the FG versus the CG simulations results in SI values of 0.58 and 0.66 for implicit and explicit solvent, respectively. This roughly good similarity between the dynamics of FG and CG simulations may suggest that both approaches sample comparable potential energy landscapes.

To exclude the possibility that the loose constraints used on the capping base pairs may generate some artifacts, we perform analogous simulations without the restraints (sys1wr and sys2wr), which give equivalent results. The only difference is the helix fraying observed at both capping base pairs of sys1wr in the ns timescale. This behavior is in agreement with FG simulations for this particular system within the simulated time [19]. However, we have also reported on the spontaneous opening and rehybridization of longer DNA filaments in the multimicro second timescale, which produce no significant changes in the global structure of the double helix [23].

3.2 Base Pair Steps and Sequence Specificity

In previous publications we have reported sequence specific effects to influence melting temperatures and breathing profiles [7,23]. To achieve a more precise evaluation of the sequence-induced structural variations we simulated systems sys5 and sys6, which contain all the unique dinucleotide base pair steps. The helical parameters are compared upon backmapping with canonical values, averaged experiments and FG simulations (Fig. 3).

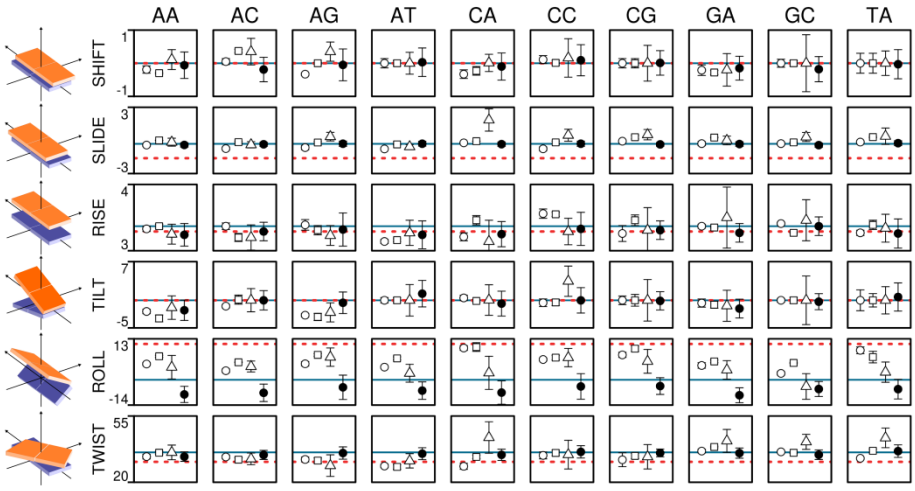


Fig. 3. Helical properties for the ten unique base pair steps. The helical properties SHIFT, SLIDE, RISE (measured in Å), TILT, ROLL and TWIST (measured in degrees) are compared for the force fields Amber99-bsc0 (empty circle), Charmm27 (squares), SIRAH (explicit solvation filled circles) and experimental x-ray measurements (triangles) including their standard deviations. All data except for that corresponding to SIRAH is taken from ref. [20]. Values of canonical B (blue line) and A (dashed red line) DNA forms are also given as references.

In general, the agreement with experiment and FG force fields is fairly good. Yet, the CG force field causes higher dispersion around average values when compared with FG simulations. The main deviation is observed for the ROLL, which shows a tendency to sample negative values. This problem is more marked for the steps CT and TC, which deviate almost 10 degrees from the canonical B-DNA. However, the impact of these deviations on the global structure and dynamics of the double helix seem to be minor, as judging from the results of the previous paragraphs.

3.3 Ionic Strength and DNA-Ion Binding

While the global distribution of cations around the DNA contributes to the stability of the double helix, the specific interaction of cations with DNA has been related with local structural distortions. In particular, the binding of sodium ions within the minor groove has been proposed to mediate its narrowing [24,25]. As quantified in our preceding publication [8], the binding of one single ion is enough to induce a sensible change in the minor groove. Increasing condensation of counterions translate in a progressively more marked minor groove narrowing.

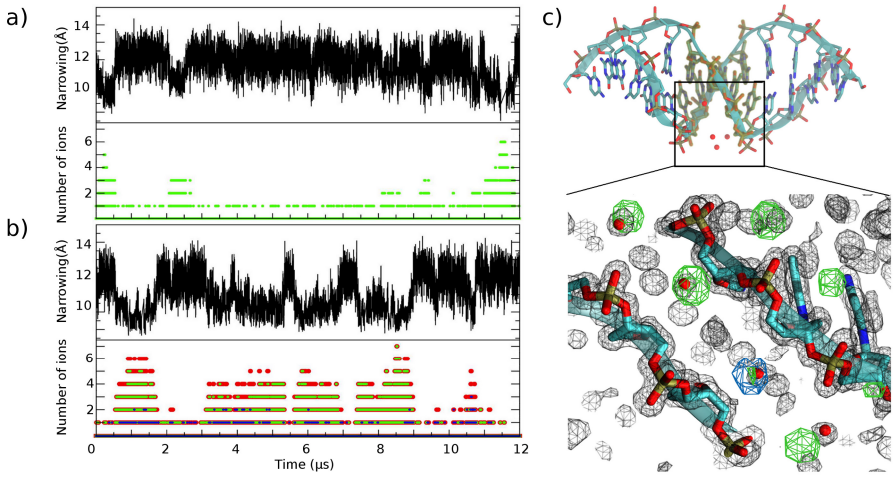


Fig. 4. Binding of cations within the minor groove of the DD dodecamer. a) Sys2wr simulation: minor groove width (top), and total number of bound cations in the minor groove as function of time. b) Idem to (a) for sys4. Green and blue dots are used for sodium and potassium, respectively. Red dots represent the sum of both electrolytes. c) Molecular representation of the DD dodecamer. The black square is zoomed in and rotated 90 degrees in the inset. Green and blue wireframes correspond to the occupational density calculated from the CG simulation for sodium and potassium ions, respectively. The gray wire frame shows the electron density from the X-ray data (PDB id: 355D [11]).

In the FG simulation (sys3) the simultaneous occupancy of the minor groove by several ions is very uncommon, but the presence of one Na^+ with residence times of 10 to 15 ns is not so rare [13]. The deformation observed in FG simulations on DNA by cations is insufficient to explain the distortions observed in the X-ray in presence

of high salt concentration [11]. Two possible causes of this disagreement may be the lack of ionic strength (sys3 contains only neutralizing counterions) or insufficient sampling. To explore this issue we perform simulation sys2wr (containing only neutralizing counterions), increasing one order of magnitude the simulation time of sys2. Moreover, we also extended to 12 μs a previously published simulation [8] of the DD dodecamer in presence of added salts (sys4, Tab. 1). The simulation of sys2wr shows that binding of one single ion is very frequent and happens in the timescale from ns to μs (Fig. 4a). Conversely, we observe very few events where several sodium ions bind simultaneously into the minor groove. When these events happen, they have a longer duration and show a correlation with the narrowing of the minor groove. Simulation of the same DNA molecule in presence of added salts (sys4) offers a complementary view of this phenomenon.

The presence of added salts increases the degree of counterions condensation around the DNA (see animation at http://www.youtube.com/watch?v=kvIWQE8UcHo&feature=youtu.be). Both cations (Na+ and K+) localize around Phosphate moieties with particularly longer residence times into the minor groove (see also animation at http://www.youtube.com/watch?v=TapsTGisEew&feature=youtu.be). From a comparison between Figs. 4a and 4b it is clear that increasing the ionic strength in the solution changes sensibly the ionic binding profile presenting microseconds-long condensation events with the simultaneous binding of up to six counterions of both species present in the solution. Notably, the average narrowing of 9.6 Å measured during long condensation events, with 3 or more bound ions, coincides precisely with X-ray determinations [11,26]. Calculation of the occupational density of ions along the MD trajectory shows that the most populated occupational sites have a rough correspondence with the geometry reported for binding sites of water, sodium and potassium within the minor groove [8]. Superposition of the occupational density with crystallographic electron density shows that this agreement is particularly good for the atoms located between the phosphate moieties (Fig. 4c). This suggests that an extended counterion condensation is needed to generate a significant and sustained bending of the DNA [27]. Measuring the total bend with the program Curves+ [21] results in an average value of 26 degrees with extreme values ranging from 10 to 50 degrees. Using implicit solvent simulations with SIRAH, we recently reported that thermal oscillations lead to DNA breathing and formation of spontaneous kinks in the double helix [23]. The DNA bending angles of nearly 80% of the known universe of protein-DNA structures (curated in the PDI database [28], http://melolab.org/pdidb/web/content/links) falls within the values sampled by our simulations. Considering the present results, we notice that 62% of all the protein-DNA complexes present a bent minor or equal to the 26 degrees induced by ion binding found in this work. This leads to the intriguing conjecture that protein-DNA recognition might exploit spontaneous fluctuations driven by the electrolytic environment. Alternatively, one might think that the concomitant binding of ions creates low entropy regions in DNA, which are more prone to be targeted by protein ligands, considerably decreasing the free energy of binding.

4 Conclusions

We presented a systematic comparison of the performance of the SIRAH CG model for DNA in implicit and explicit solvation against FG simulations and experimental

data. It turns out that both approximations provide a good description of the structural and dynamical features of DNA. The gross determinants of the structural stability of the double helix suggest that, upon backmapping, the information obtained from the CG simulations is almost as accurate as that provided by FG techniques. The specific and reversible binding of counterions within the minor groove generates microseconds-long narrowing, that probably relates to bent DNA conformations up to ~50 degrees. This distortion translated in a narrowing of the minor groove, which may be stable in the multi-microsecond time window.

The agreement with crystallographic data [26] and the comparison with all the high resolution protein-DNA complexes reported in the PDB provides a validation for our model and highlights the importance of the proper treatment of ionic strength effects. CG simulations sample a large range of DNA bending conformations seen in protein-DNA crystals, suggesting that, thermally induced oscillations of naked DNA encompasses the distortions required for protein binding. In line with similar conclusions conducted at the FG level [29], this add a new piece of evidence in favor of the prevalence of 'conformational selection' versus 'induced fit' paradigms.

The simulation schemes presented here result in a speed up respect to their corresponding FG simulations of 800 and 2400 times for explicit and implicit solvent simulations, respectively. Since the high computational cost associated to atomistic simulations precludes the study of many interesting phenomena, this kind of approaches is expected to become of regularly use and interest for the broad scientific community.

Acknowledgements. This work was supported by ANII – Agencia Nacional de Investigación e Innovación, Programa de Apoyo Sectorial a la Estrategia Nacional de Innovación – INNOVA URUGUAY (Agreement n8 DCI – ALA / 2007 / 19.040 between Uruguay and the European Commission). A.B. is beneficiary of the National Fellowship System of ANII. P.D.D. and S.P. appreciate support from the National Scientific Program of ANII (SNI) and PEDECIBA. We thank Adrian Roitberg for useful discussions about the implementation of the WT4 topology in the AMBER package and Sebastian Ferreira for excellent technical support.

References

[1] Karplus, M., McCammon, J.A.: Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. 9, 646–652 (2002)

[2] Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science 330, 341–346 (2010)

[3] Arkhipov, A., Yin, Y., Schulten, K.: Four-scale description of membrane sculpting by BAR domains. Biophys. J. 95, 2806–2821 (2008)

[4] Yin, Y., Arkhipov, A., Schulten, K.: Simulations of membrane tubulation by lattices of amphiphysin N-BAR domains. Structure 17, 882–892 (2009)

[5] Voth, G.A.: Coarse-Graining of Condensed Phase and Biomolecular Systems. Taylor & Francis Group, New-York (2009)

[6] Potoyan, D., Savelyev, A., Papoian, G.: Recent successes in coarse-grained modeling of DNA. WIREs Comput. Mol. Sci. 3, 69–83 (2013)

[7] Dans, P.D., Zeida, A., Machado, M.R., Pantano, S.: A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics. J. Chem. Theory Comput. 6, 1711–1725 (2010)
[8] Darré, L., Machado, M.R., Dans, P.D., Herrera, F.E., Pantano, S.: Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? J. Chem. Theory Comput. 6, 3793–3807 (2010)
[9] Machado, M.R., Dans, P.D., Pantano, S.: A hybrid all-atom/coarse grain model for multiscale simulations of DNA. Phys. Chem. Chem. Phys. 13, 18134–18144 (2011)
[10] Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Itakura, K., Dickerson, R.E.: Structure of a B-DNA dodecamer: conformation and dynamics. Proc. Natl. Acad. Sci. U. S. A. 78, 2179–2183 (1981)
[11] Shui, X., McFail-Isom, L., Hu, G.G., Williams, L.D.: The B-DNA dodecamer at high resolution reveals a spine of water on sodium. Biochemistry 37, 8341–8355 (1998)
[12] Hawkins, G.D., Cramer, C.J., Truhlar, D.G.: Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J. Phys. Chem. 100, 19839 (1996)
[13] Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham III, T.E., Laughton, C.A., Orozco, M.: Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys. J. 92, 3817–3829 (2007)
[14] Pastor, R.W., Brooks, B.R., Szabo, A.: An analysis of the accuracy of Langevin and molecular dynamics algorithms. Mol. Phys. 65, 1409–1419 (1988)
[15] Wu, X., Brooks, B.R.: Self-guided Langevin dynamics simulation method. Chem. Phys. Lett. 381, 512–518 (2003)
[16] Hess, B., Kutzner, C., van de Spoel, D., Lindahl, E.: GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J. Chem. Theo. Comp. 4, 435–447 (2008)
[17] Essmann, U., Perera, L., Berkowitz, M.L., Darden, T.A., Lee, H., Pedersen, L.: A smooth particle mesh ewald potential. J. Chem. Phys. 103, 8577–8592 (1995)
[18] Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., Haak, J.R.: Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81, 3684–3691 (1984)
[19] Pérez, A., Luque, F.J., Orozco, M.: Dynamics of B-DNA on the Microsecond Time Scale. J. Am. Chem. Soc. 129, 14739–14745 (2007)
[20] Perez, A., Lankas, F., Luque, F.J., Orozco, M.: Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Res. 36, 2379–2394 (2008)
[21] Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D., Zakrzewska, K.: Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res. 37, 5917–5929 (2009)
[22] Humphrey, W., Dalke, A., Schulten, K.: VMD - Visual Molecular Dynamics. J. Molec. Graphics 14, 33–38 (1996)
[23] Zeida, A., Machado, M.R., Dans, P.D., Pantano, S.: Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. 86, 021903 (2012)
[24] Hamelberg, D., Williams, L.D., Wilson, W.D.: Influence of the dynamic positions of cations on the structure of the DNA minor groove: sequence-dependent effects. J. Am. Chem. Soc. 123, 7745–7755 (2001)
[25] Hamelberg, D., Williams, L.D., Wilson, W.D.: Effect of a neutralized phosphate backbone on the minor groove of B-DNA: molecular dynamics simulation studies. Nucleic Acids Res. 30, 3615–3623 (2002)

[26] Shui, X., Sines, C.C., McFail-Isom, L., VanDerveer, D., Williams, L.D.: Structure of the potassium form of CGCGAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. Biochemistry 37, 16877–16887 (1998)
[27] Spiriti, J., Kamberaj, H., de Graff, A., Thorpe, M.F., van der Vaart, A.: DNA Bending through Large Angles Is Aided by Ionic Screening. J. Chem. Theo. Comp. 8, 2145–2156 (2012)
[28] Norambuena, T., Melo, F.: The Protein-DNA Interface database. BMC Bioinformatics 11, 262 (2010)
[29] Dans, P.D., Perez, A., Faustino, I., Lavery, R., Orozco, M.: Exploring polymorphisms in B-DNA helical conformations. Nucleic Acids Res. 40, 10668–10678 (2012)

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

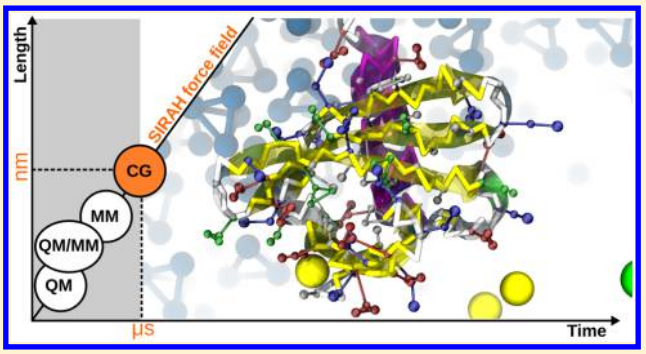
SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics

Leonardo Darré,[†,‡] Matías Rodrigo Machado,[†] Astrid Febe Brandner,[†] Humberto Carlos González,[†] Sebastián Ferreira,[†] and Sergio Pantano*[,†]

[†]Institut Pasteur de Montevideo, Montevideo, Uruguay

[‡]Department of Chemistry, King's College, London, United Kingdom

ABSTRACT: Modeling of macromolecular structures and interactions represents an important challenge for computational biology, involving different time and length scales. However, this task can be facilitated through the use of coarse-grained (CG) models, which reduce the number of degrees of freedom and allow efficient exploration of complex conformational spaces. This article presents a new CG protein model named SIRAH, developed to work with explicit solvent and to capture sequence, temperature, and ionic strength effects in a topologically unbiased manner. SIRAH is implemented in GROMACS, and interactions are calculated using a standard pairwise Hamiltonian for classical molecular dynamics simulations. We present a set of simulations that test the capability of SIRAH to produce a qualitatively correct solvation on different amino acids, hydrophilic/hydrophobic interactions, and long-range electrostatic recognition leading to spontaneous association of unstructured peptides and stable structures of single polypeptides and protein−protein complexes.



INTRODUCTION

The exponential growth of computer power added to the development of faster algorithms has contributed to make molecular simulations a reliable tool for the study of biomolecular systems. Nevertheless, direct comparison with experimental data is often difficult owing to the large size and long time scales needed for a proper description of the complex biological environment. These difficulties have motivated the development of simplified methods aimed to bridge the gap between experiments and simulations. A large number of coarse-grained (CG) molecular representations have been described in the literature for the simulation of the most common biological species.[1−18] In general, the microscopic details are coarsened following either top-down or bottom-up approaches. In bottom-up schemes, a given Hamiltonian function is chosen and parametrized to fit fine-grained (FG) simulations taken as a reference. Several strategies to derive CG potentials have been developed on the basis of mining degrees of freedom from FG simulations through force matching techniques, Boltzmann inversion, thermodynamic integration, etc.[19,20] In top-down approaches, force fields are often tailored on the basis of physicochemical intuition and/or trial and error simulations, and interaction parameters are fitted to match available experimental data.

Bottom-up strategies can produce very accurate potentials and are very well suited for the description of uniform systems. However, it may be difficult to derive a general and transferable CG force field for highly heterogeneous macromolecules as proteins.[20] On the other hand, the accuracy of top-down models may be strongly related to the availability of experimental data but may provide potentials that are more easily transferable.[21] For recent reviews on different CG approaches, see Ingolfsson et al.[22] and Brini et al.[23]

Recently, our group has undertaken the initiative to develop a CG force field for biomolecules named SIRAH (http://www.sirahff.com). We followed a top-down approach fitting structural properties of macromolecules using a standard pairwise Hamiltonian common to most MD simulation packages. So far, the SIRAH force field includes parameters and topologies for simulating DNA using an implicit solvation scheme[24,25] or embedded in an explicit CG representation of aqueous solvation.[26] Our CG model for water (named WatFour or WT4 for shortness) is composed by four linked beads, each carrying a partial charge. This confers to WT4 the capacity to create its own dielectric permittivity, while the use of CG electrolytes helps to account for ionic strength effects and osmotic pressure.[26] The WT4 model has been recently shown to be suitable for hybrid or dual-resolution simulations, where regions of interest within molecular systems can be treated at full atomic detail, while bulk regions of the solvent are simulated at the CG level without perturbing the structure and dynamics of the atomistic part.[27−29] Along this line, we have also expanded our force field to consider a dual-resolution version of double stranded DNA[30] compatible with the FG AMBER99 force field.[31]

Received: August 26, 2014

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Here, we present a novel CG model for proteins and peptides, which works in combination with the WT4 model for explicit solvation. The set of parameters presented here aims to overcome some common limitations of CG force fields as the use of uniform dielectric constant, lack of long-range interactions, use of topological information to maintain the secondary structure, implicit or no ionic strength effects, etc. The performance of our model is illustrated on a series of simulations on different peptides, proteins, and protein−protein complexes chosen as working examples to highlight particular characteristics of the force field. These examples include sequence, temperature, and ionic strength-dependent conformational changes of helical peptides, spontaneous formation β-strands, reproduction of stable structures of proteins with different folds, and protein−protein complexes.

Derivation of the Model. The conception of the protein model follows the general structure-based philosophy used for DNA and aqueous solvent. The FG-to-CG mapping scheme is based on physicochemical intuition and uses the position of real atoms to place CG beads. Interaction parameters are empirically fitted to reproduce structural features of target molecules. In particular, the internal coordinates of our tetrahedral CG water model are based on the transient configuration of water clusters, and interaction parameters between water and free electrolytes are fitted to match the structure of the second solvation shell as determined by neutron scattering experiments.[26] Similarly, our CG model of DNA contains two beads at the position of the phosphorus and C5′ carbon, representing the 5′−3′ polarity of the backbone. A point charge of $-1e$ is placed on the phosphate, and the sizes of the beads represent the excluded volume of their atomistic counterpart.[24] In contrast, the beads representing the base moiety present an atomistic size to allow for proper stacking between bases and hydrogen bond-like interactions in the Watson−Crick region. This nonuniform granularity is used to represent nearly atomistic interactions at the bases (i.e., hydrogen bonds and stacking) and coarser (less specific) interactions at the backbone. It is worth noting that this mapping combined with a uniform mass distribution, bond and angular stretching constants leads to nearly atomistic reproduction of DNA structure and dynamics, including ionic strength conformational effects.[32] Fitting of the WT4 bead's mass to water density resulted in 50 au, which is very close to the value obtained by summing the FG masses of the four nucleotides and taking the average on the six CG beads composing one of them (i.e., 51.53 au[26]). Therefore, in this version of the force field, we decided to adopt a uniform value of 50 au for all the protein beads. This choice is also computationally convenient because in combination with bond-stretching parameters it allows for time steps of 20 fs.

For the CG mapping of proteins, the peptide bonds are treated with a relatively low granularity, keeping the positions of the nitrogen (N), α carbon (Cα), and oxygen (O) (Figure 1A), while side chains are modeled at lower degree of detail (see below). Similar strategies have been previously presented, as for instance refs 10, 11, and 33. This choice might be advantageous in relation to more uniform CG mappings presented in the literature. In particular, the correspondence between FG and CG conformational space is univocal, leading to a straightforward physical interpretation of the conformational space explored by peptides/proteins. Additionally, backbone representations considering only one bead on the Cα may need specific constraints to reproduce secondary

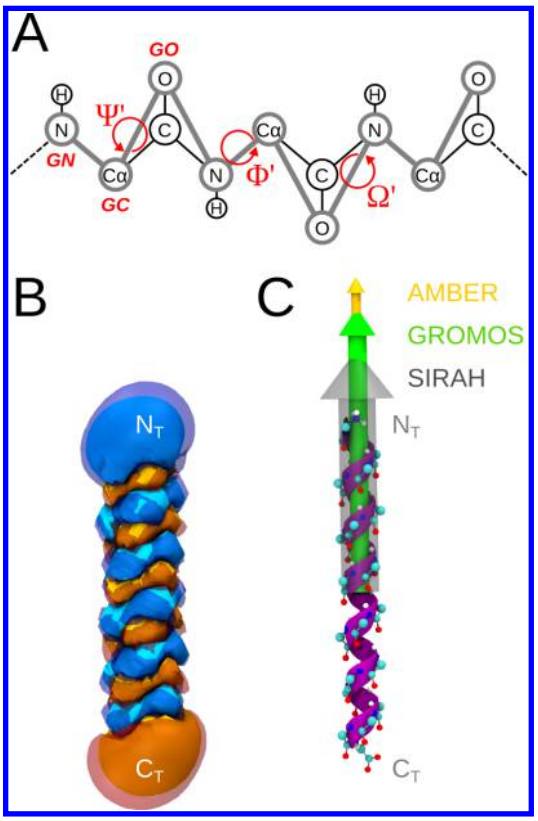


Figure 1. Structure and electrostatics of the SIRAH protein backbone. (A) CG mapping scheme. Gray circles indicate the position of the atoms used to define the CG beads in the backbone, while thick gray lines indicate their connectivity. The main dihedral angles at the CG level are shown. (B) Comparison between electrostatic profiles in the neighborhood of a α-helix generated by SIRAH and AMBER99. Solid and semi-transparent surfaces correspond to AMBER99 and SIRAH, respectively. Positive (blue) and negative (red/orange) isosurfaces are traced at ±10 mV. (C) Dipole moments calculated on the same helix using the charge distribution of AMBER99 (yellow), Gromos96/45a3 (green), and SIRAH (gray, semi-transparent). The origin of the dipoles is displaced in the vertical direction to the center of the helix for visualization purposes.

structure elements depending on the peptide conformation.[34] In contrast, our choice of backbone beads with atomistic sizes is conformation independent and allows for the formation of the optimally compacted topology of α-helices.[35] Moreover, the use of partial charges on each bead can roughly account for the formation of hydrogen bond-like interactions, stabilizing the formation of the α-helices and β-sheets without imposing ad hoc constraints. On the other hand, this mapping based on effective interaction points within functional groups may not be easy to generalize. Hence, simple mapping procedures as the "four heavy atoms to one CG bead" rule used by the popular Martini force field[9] may be difficult to derive in our case.

The connectivity between N, Cα, and O beads (named GN, GC, and GO, respectively) is presented in Figure 1A. Equilibrium distances and angles for bonded parameters between all backbone beads are adopted from the minimum energy conformation of FG glycine tripeptides using the AMBER99 force field.[31] Notice that these quantities are independent of the conformation or the entity of the amino acid used and the actual connectivity in the FG systems. As an initial guess, we used the same bond and angular stretching force constants used for the DNA model, i.e., 41840 kJ/mol

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

nm^2 and 627.6 kJ/mol rad^2. As shown in the Results section, this choice leads to stable MD trajectories of peptides and proteins without further refinement.

A combination of dihedral angles is imposed to the quadruplets defined by the backbone beads (Figure 1A and Table 1). This representation keeps a correspondence between

Table 1. Torsional Parameters in SIRAH Force Field for Proteins[a]

	m	k_m (kJ/mol)	n_m	α_{0m} (deg)
	backbone			
Ψ'(GN$_i$-GC$_i$-GO$_i$-GN$_{i+1}$)	1	1.8	1	160
	2	15.0	2	−270
	3	4.8	1	−130
Φ'(GO$_{i-1}$-GN$_i$-GC$_i$-GO$_i$)	1	11.5	4	60
	2	22.4	3	60
	3	5.1	4	180
	4	16.0	5	−90
Ω'(GC$_{i-1}$-GO$_i$-GN$_i$-GC$_i$)	1	60.0	1	0
	amino acid "L" chirality			
side-chain-GC-GN-GO[b]	1	100	1	35
	tryptophan side chain (improper)			
BPE-BCG-BCZ-BCE[c]	1	250.0	1	180
BCZ-BCG-BPE-BNE[c]	1	100.0	1	0

[a]The torsional potential in GROMACS is defined as $V = \sum_m k_m [1 + \cos(n_m\alpha + \alpha_{0m})]$. [b]Applies to the first bead in the side chain. [c]See Figure 2 for naming.

FG and CG dihedral angles Ψ, Φ, and Ω, defining the secondary structure (see Calculated Properties). Hence, we set these dihedral angles by a polynomial fitting forcing the existence of minima in the two more stable conformations, i.e., α-helices (Ψ = −57°, Ω = −47°) and β-strands (Ψ = 150°, Ω = −80°). The suitability of the functions obtained was initially tested on peptides in α-helical and β-sheet conformations and further refined on alpha beta (a + b) proteins. The final set of functions containing three and four terms for Ψ and Φ, respectively, is shown in Table 1. At this stage of development, only peptides in *trans* configuration are considered. Amino acids in *cis* conformations, which are much less frequent in nature, will be incorporated in further versions of the force field. Hence, Ω is represented with a single cosine function with one minimum in the *trans* region (Table 1).

The partial charges on the backbone's beads are set to roughly reproduce the electrostatic potential generated by a fully atomistic α-helix, also known as the helix macrodipole. Neutral and charged termini are available in the current version of SIRAH. Charged termini are simply constructed by adding a ± 1*e* charge on the N and/or O terminal beads of the chain, respectively.

To compare the similarity between the electrostatic potentials generated by popular FG force fields and SIRAH, we considered a 24 residues long polyglycine in canonical α-helical conformation. This helical extension was chosen, as it is the approximate length of a typical transmembrane helix. The electrostatic potential is calculated by solving the linearized Poisson−Boltzmann equation on a box with 12 nm edges and grid spacing of 0.05 nm imposing a zero value for the potential at the borders using APBS.[36] The qualitative likeliness between the electrostatic potential generated by SIRAH and AMBER99 can be observed in Figure 1B. A more quantitative evaluation can be acquired by comparing the dipole moments generated by different force fields (Figure 1C). Considering neutral terminals results in nearly collinear vectors with modules of 65 D, 76.5 D, and 84.6 D for SIRAH, Gromos96/45a3, and AMBER99, respectively. If zwitterionic terminals are used, we obtain dipole moments of 285.1 D, 267.3 D, and 259.5 D for SIRAH, Gromos96/45a3, and AMBER99, respectively. This suggests that long-range interactions reproduced by our CG scheme at the backbone level are comparable to those of popular FG force fields.

Finally, van der Waals (vdW) interactions within backbone beads are set to the same values of AMBER99, ensuring a correct degree of compaction upon formation of α-helices and β-sheets.

The CG topology of the side chains follows the philosophy of representing interaction points according to the characteristics of each residue (Figure 2). Hydrophobic residues (Val, Ile, Leu, and Met) are represented by one single bead. Aromatic side chains are mapped to three (Phe, His, and Tyr) or five beads on a plane (Trp). The beads of polar and charged side chains are mapped on charged groups or acceptors/donors of hydrogen bonds.

Following the same procedure described for the backbone, bond and angular equilibrium positions are taken from the AMBER99 force field, while DNA values are used as initial guess of force constants. While these values are well suited for the backbone beads, iterative sampling revealed that the values reported in Table 2 perform better for side chains. Improper dihedral angles are imposed on the backbone and side-chain beads to ensure the "L" chirality of the amino acids. Additionally, two improper dihedrals are used on the side chain of tryptophan to force its planarity (Table 1).

Point charges are assigned under the general hypothesis that functional groups bearing more hydrogen bond acceptors/donors at the FG level should carry higher charges in the CG scheme to establish stronger electrostatic interactions. In this line, hydrophobic beads carry a zero charge, and polar moieties within aromatic side chains carry a charge of ±0.1*e*. Hydroxyl and amide groups are charged ±0.2 *e* and ±0.4 *e*, respectively, and charged amino acids present a unitary charge spread on the side-chain beads (Figure 2). In the present version, all protonation states are considered at pH 7.

The vdW parameters are treated as free variables to modulate interactions. As an initial guess, we started with sigma and epsilon corresponding to those of the WT4 beads (Figure 2). The values of the radii (sigma) underwent a progressive adjustment on static structures until no steric clashes are present. This is performed by varying the radii to obtain nonpositive values when calculating the vdW component of the potential energy pairwise on each couple of amino acids. After setting the sigma values on static structures, the epsilon parameters are iteratively varied to obtain stable trajectories. Aimed to limit the search of optimal combinations, these parameters are assigned by residue type (hydrophobic, aromatic, etc.). Ad hoc modifications are introduced to single beads of proline, methionine, lysine, and alanine and CG beads mapped on hydrogen atoms (Figure 2).

van der Waals interactions are, in general, calculated according to the Lorentz−Berthelot combination rules. However, owing to the high granularity of the CG solvent, it may not properly intercalate between polar moieties, introducing spurious conformational effects. Similarly, we found that the atomistic parameters used for backbone−backbone interactions result in poorly balanced interaction with

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

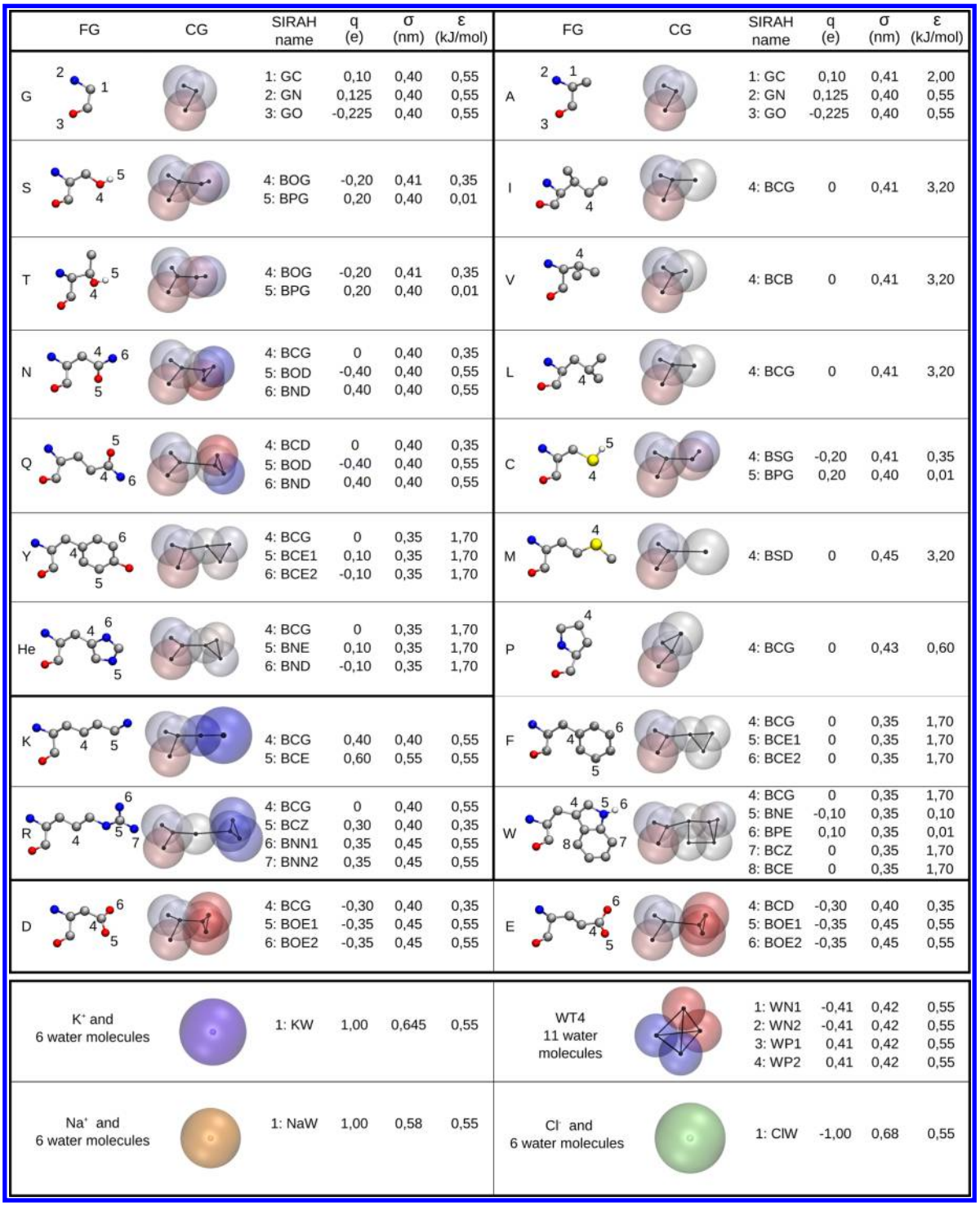


Figure 2. SIRAH representation of amino acids and solvent. The one-letter code, FG heavy atoms, CG representation, bead's names, partial charges, and vdW parameters are presented for each amino acid. Only hydrogen atoms used for the CG mapping are shown. Numbers near the FG atoms indicate the position for the corresponding CG beads. Backbone mapping is indicated only for glycine and alanine. FG atoms are colored by name, while CG beads of amino acid and WT4 are colored by charge range (negative, red; positive, blue). The sizes of all CG beads are at scale and correspond to their actual vdW radii.

CG beads. To ameliorate this flaw, we set few specific corrections to the vdW interactions that avoid over stabilization of salt bridges, hydrogen bond-like interactions, or spurious protein−ion contacts (Figure 3). These corrections have been added case-by-case following an iterative trial and error procedure.

A comprehensive view of the topologies, sizes, names, charges, and vdW parameters of the 20 amino acids, water, and electrolytes available in the SIRAH force field is presented in Figure 2.

Computational Details. The simulation protocols applied in this work are essentially the same used for any plain molecular dynamics (MD) simulation: (i) Initial coordinates

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 2. Force Constants of Angular Parameters[a]

position of the beads	k (kJ/mol rad^2)	applies to
backbone−backbone−backbone	627.6	all residues
backbone (terminal)−backbone−side-chain	60.0	all terminal residues
backbone−backbone−side-chain (not aromatic)	100.0	Ser, Thr, Asn, Gln, Lys, Arg, Asp, Glu, Ile, Leu, Val, Met (not terminal)
backbone−backbone−side-chain (aromatic)	150.0	Tyr, His, Phe, Trp (not terminal)
backbone−side-chain−side-chain	50.0	Ser, Thr, Cys, His, Tyr, Phe, Trp
backbone−side-chain−side-chain	10.0	Arg, Lys, Asn, Gln, Asp, Glu,
side-chain−side-chain−side-chain	0.0	Asp, Gln, Tyr, His, Arg, Asp, Glu, Phe, Trp

[a]Equilibrium values are taken from AMBER99. The angular potential in GROMACS is defined as $V = \mathbf{k}/2(\sigma-\sigma_0)^2$.

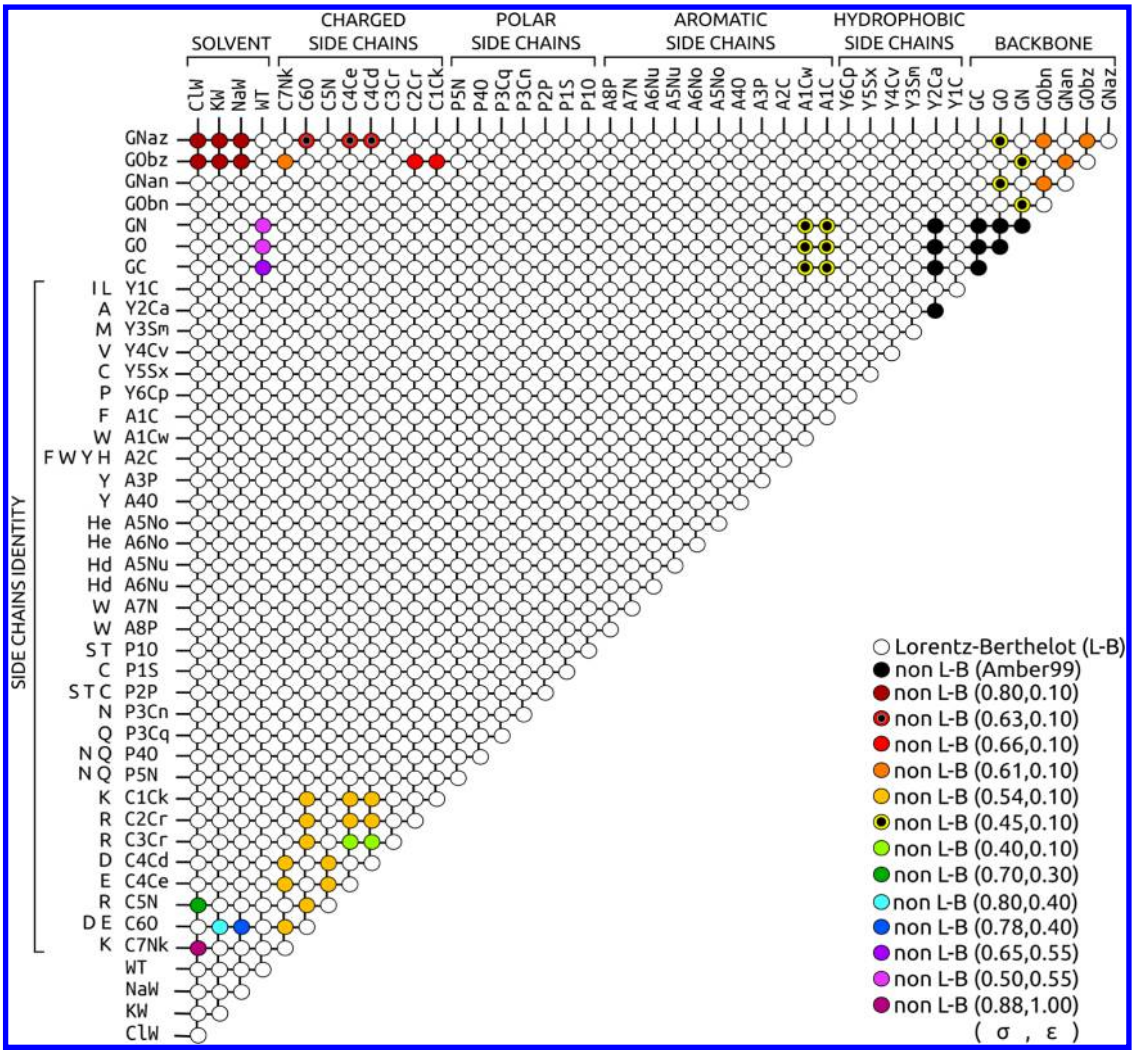


Figure 3. Scheme of vdW interactions in SIRAH. All the atom types currently defined in the SIRAH force field for proteins, ions, and water are arranged in a diagonal matrix. Interactions calculated with the Lorentz−Berthelot combination rules are indicated with empty circles. Colored circles correspond to interactions computed outside Lorentz−Berthelot combination rules using specific values indicated in the legend.

are taken from the PDB and protonated at pH 7. (ii) The FG structures are converted to CG by an ad hoc script. (iii) The CG model is solvated using a prestabilized box of WT4 molecules. (iv) Ionic strength or electroneutrality is set by substituting WT4 molecules by electrolytes (CG versions of Na^+, K^+, and Cl^- are currently available). As the densities of WT4 and FG water are coincident, and considering the molecular weight of WT4 (200 g/mol), a proportion of one ionic couple per 50 WT4 molecules roughly represents a 0.1 M concentration. The simulation protocols also include (v) energy minimization, (vi) equilibration MD, and (vii) production run.

Starting structures are protonated using pdb2pqr,[37] converted to CG and solvated in an octahedral box with a solute−box distance of 1.5 nm. The ionic strength and target temperature are set to values reported in each case. Simulations are performed using GROMACS 4.5.5 (http://www.gromacs.org) with a time step of 20 fs updating the neighbor list every 10 steps. Electrostatic interactions are calculated using Particle Mesh Ewald (PME)[38] with a direct cut off of 1.2 nm and a grid spacing of 0.2 nm. The same cut off is used for vdW interactions. Energy minimization is carried out by 1000 iterations of the steepest descent algorithm. Equilibration

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

dynamics is accomplished by 5 ns of MD with positional restraints of 1000 kJ/mol nm^2 applied to all the protein beads. Production runs are performed by 1 μs (unless otherwise stated) in the NPT ensemble using v-rescale thermostat[39] and Parrinello–Rahman barostat.[40]

Calculated Properties.

(1) Hydrogen bonds (H-bonds) are operationally defined to be formed if two beads of opposite charge are at a distance ≤0.4 nm. Similarly, a salt bridge is considered to exist if a couple of beads of opposite charge belonging to ionic residues are within a distance ≤0.6 nm.

(2) Secondary structure is calculated as a function of the dihedral angles along the backbone beads and defined between ±180°. Owing to the absence of the carbonyl carbon in the CG backbone, the φ and ψ dihedral angles defining the Ramachandran plot are calculated as (see Figure 2 for naming)

$$\Phi = \Phi'(GO_{i-1} - GN_i - GC_i - GO_i) - 20°$$

$$\Psi = \Psi'(GN_i - GC_i - GO_i - GN_{i+1}) + 13°$$

From this definition the conformation of each residue is assigned to the categories: helical (H), extended (E), or coil (C) using the following criteria

$$SS_i(\Phi, \Psi): \begin{cases} \text{H: } -180° \le \Phi \le 10° \text{ and } -120° \le \Psi \le 45° \\ \text{E: } (-180° \le \Phi \le 0° \text{ or } \Phi > 135°) \text{ and } (-180° \le \Psi \le -120° \text{ or } 45° \le \Psi \le 180°) \\ \text{C: otherwise} \end{cases}$$

where ss_i is the conformation assigned to residue i. Additionally, for residues to belong to the H category, either the GO or the GN beads must be within the H-bond definition with a residue separated by four consecutive positions in the chain. Similarly, for residues to be considered in E conformation, GO, GN, or GC beads must form H-bonds with backbone neighbors separated by more than two consecutive positions in the chain and belonging to the same category.

Residues not assigned to H or E are considered, for the sake of simplicity, as members of category C. Notice that the secondary structure is dynamic and can change during the trajectory.

(3) Contacts between residues are considered to exist if two Cα beads are within 0.8 nm. Native contacts are defined on the experimental structure, and its conservation is reported as a percentage along the trajectory. Global contacts are calculated on all the residues in the chain, while noncontiguous contacts are computed excluding residue pairs within five neighboring positions in sequence. The accuracy is evaluated as Acc = TP/(TP + FP), where TP (true positives) is the number of correctly reproduced contacts, and FP (false positives) are new contacts during the simulation.

(4) Peptide aggregates are considered to exist only if the distance between centers of mass of amino acids is within 0.5 nm, otherwise peptides are considered to belong to different aggregates. GROMACS' g_clustersize tool is used to analyze the coordinates of the peptides every 1 ns during 5 μs (a total of 5000 simulation frames).

(5) Protein accessible surfaces (SAS) are calculated with GROMACS' g_sas with a probe radius of 0.21 nm (i.e., the radius of a WT4 bead). Beads are considered to be hydrophobic if they carry a zero charge. Notice that specific vdW radii of SIRAH reported in Figure 2 have to be used.

(6) Protein–protein interfaces are calculated as $0.5(SAS_{Prot1} + SAS_{Prot2} - SAS_{Complex})$

(7) RMSD, B-factors, and gyration radii are calculated on the Cα beads using the standard GROMACS' tools.

RESULTS AND DISCUSSION

In this section, we test our CG model on a series of peptides and protein systems. The main goal of this set of simulations is not that of furnishing new biophysical insights for each particular system but highlighting specific features of our force field in different possible scenarios.

Test Case 1: Cytoplasmic Domain of Phospholamban. Phospholamban (PLB) is a small membrane protein that regulates the activity of the calcium ATP-ase in the cardiac, slow-twitch, and smooth muscle sarcoplasmic reticulum.[41] It consists of a cytoplasmic moiety subdivided in domains Ia (residues 1–17), Ib (residues 18–30), and transmembrane (residues 31–52). Several NMR studies have shown that domain Ia adopts an α-helical configuration, independent of the rest of the protein.[42,43] Early attempts for structural determination also demonstrated the importance of added salts in the solution to stabilize the structure of this protein segment.[42]

The small size of this domain and the availability of structural experimental data make it suitable as an initial test case for our force field on a biologically relevant system. Moreover, the simplicity of this motif and the presence of polar, hydrophobic, and charged amino acids, all exposed to the solvent (Figure 4A), allows us to illustrate characteristic protein–solvent interactions disregarding the structural context.

Protein–Solvent Interactions in Peptide Context. We first analyze the solvation structure around amino acids with different physicochemical characteristics in domain Ia of PLB. With this aim, we take the first 17 amino acids of the structure 1FJK (Figure 4A) and map them into our CG representation (Figure 4B). The CG peptide is then embedded in a solvation box containing 287 WT4 molecules and 9 and 13 CG ions (NaW+ and ClW–, respectively), which ensure electroneutrality and represent a salt concentration of ∼0.2 M (Figure 4C). After energy minimization and equilibration, MD simulations are carried out at 300 K and 1 atm for 1 μs.

The collected data is first used to compute the solvation structure around different moieties in terms of the radial distribution function (RDF) of water. Calculation of solvent's RDF around the backbone increases smoothly after 0.5 nm (Figure 5A) as a consequence of specific backbone–WT4 vdW interactions (Figure 3). Consequently, the solvent distribution

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

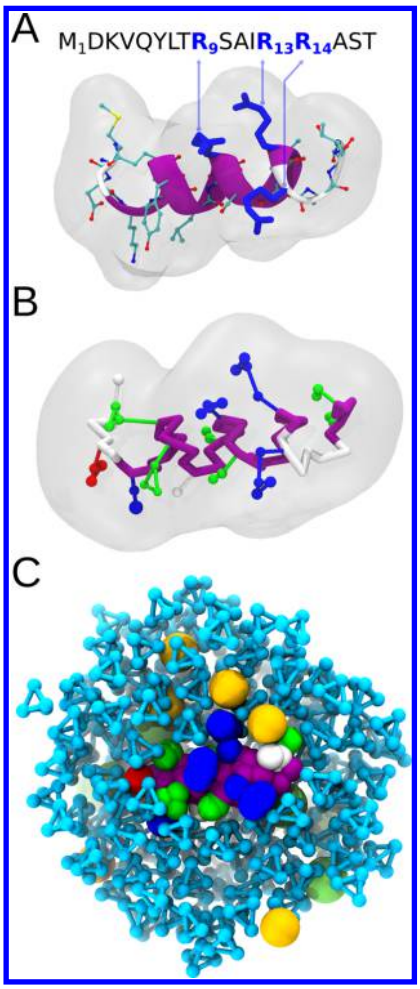


Figure 4. Molecular representations of domain Ia of PLB. (A) Sequence and starting MD conformer corresponding to the first 17 residues of the NMR structure 1FJK. The cartoon traced on the backbone atoms indicates the secondary structure (helix, purple; coil, white). Heavy atoms are colored by element, except for Arg9, 13, and 14, which are shown in blue. The semi-transparent surface corresponds to the SAS traced with a probe radius of 0.14 nm. (B) SIRAH mapping of the peptide presented in panel A. Amino acids are colored by residue type (positive, blue; negative, red; polar, green; and hydrophobic, white). In this case, the solvent-accessible surface is traced with a probe radius of 0.21 nm, which corresponds to the dimension of a WT4 bead. (C) Typical SIRAH simulation box. The domain Ia of PLB is placed within an octahedron box and solvated with WT4 and CG salt. The sizes of the solute (colored as in panel B) and ions (NaW, yellow; ClW, green) correspond to their vdW radii, while WT4 molecules (light blue tetrahedrons) are shown with balls and sticks for visualization sake.

around the main chain results in a hydrophobic-like profile. Hence, the global shape of this distribution function is roughly comparable to that measured around the hydrophobic side chain of Ile12, which is taken as a representative hydrophobic bead (Figure 5A). Discriminating between negative or positive beads of WT4 (Figure 2) does not significantly affect the RDFs neither for the Cα bead nor for the hydrophobic residues (not shown).

In stark contrast, the solvent organization around charged side chains presents qualitative and quantitatively different profiles. The organization of the positive beads of WT4 around BOE1 in D2 (see Figure 2 for naming) shows the typical shape of a hydrophilic profile (Figure 5B). We found a marked peak at 0.42 nm, followed by a minimum, indicating the passage from a region of high population of positive beads to a depletion caused by electrostatic interactions. In agreement with this electrostatics-driven solvent organization, the distribution of positive WT4 beads around BCE at K3 is shifted to the right and follows an opposite trend to that shown for D2 (Figure 5B). The role of electrostatics in the solvent organization is confirmed if we consider the arrangement of *negative* beads around the same moieties. The partial charge placed on the BCE bead results in a significantly sharp solvation peak at 0.45 nm, while a minimum is observed in the close neighborhood of BOE1 (Figure 5C). After nearly 1.2−1.5 nm, all the distributions flatten, converging to bulk values with oscillations presenting a periodicity corresponding to the size of WT4 molecules.[26]

As representative of a polar residue, we consider the side chain of S16. In this case, we see that the CG beads placed at the positions of the gamma oxygen (BOG) and hydrogen (BPG) of S16 generate a less marked organization of the surrounding solvent than charged amino acids, being actually more similar to the hydrophobic moiety of I12 (Figure 5). The impaired capacity of serine to induce a marked hydrophilic-like solvation structure can be ascribed to the small charges placed on this side chain and to the separation between the beads (i.e., bond distance in a hydroxyl group). However, even these small partial charges are able to create differences in solvent organization. This is more evident in the distribution of negative beads, which exhibits a right-shift for the BOG bead (Figure 5C).

Sequence, Temperature, and Ionic Strength Conformational Effects. A further step in assessing the performance of our force field is to explore the conformational response of SIRAH to variations introduced in the molecular system.

To define a reference state, we briefly describe the results of the simulation introduced in the previous section. Along the trajectory, M1, D2, S16, and T17 very frequently visit disordered conformations. However, these transitions do not compromise the global structural stability of the peptide within the μs time scale. Indeed, the RMSD oscillates around 0.27 nm with a mean helical content of 73%. These slight conformational changes are in good agreement with the variation observed in the family of conformers of the cytoplasmic domain of PLB reported on the basis of NMR studies[42] and previous FG simulations on a shorter time scale.[44]

On this simple system, we sought to analyze the response of SIRAH to variations in (i) sequence, (ii) temperature, and (iii) ionic strength.

(i) To rule out possible conformational biases in the parametrization of the backbone beads favoring secondary structures, we used the same starting conformation of PLB but mutating all amino acids to glycine. As shown in Figure 5D, the helical content of the polyglycine falls rapidly arriving to nearly 50% already within the first 2 ns and continues to decrease until the 6 ns to then oscillate between 0% and 17%. These changes are accompanied by an increase in the coil content and to a minor extent the extended configurations. Conformational transitions may happen one residue at a time or more than one simultaneously. These transitions, which imply the rupture of H-bond interactions between backbone beads and rotations along dihedral angles, happen

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

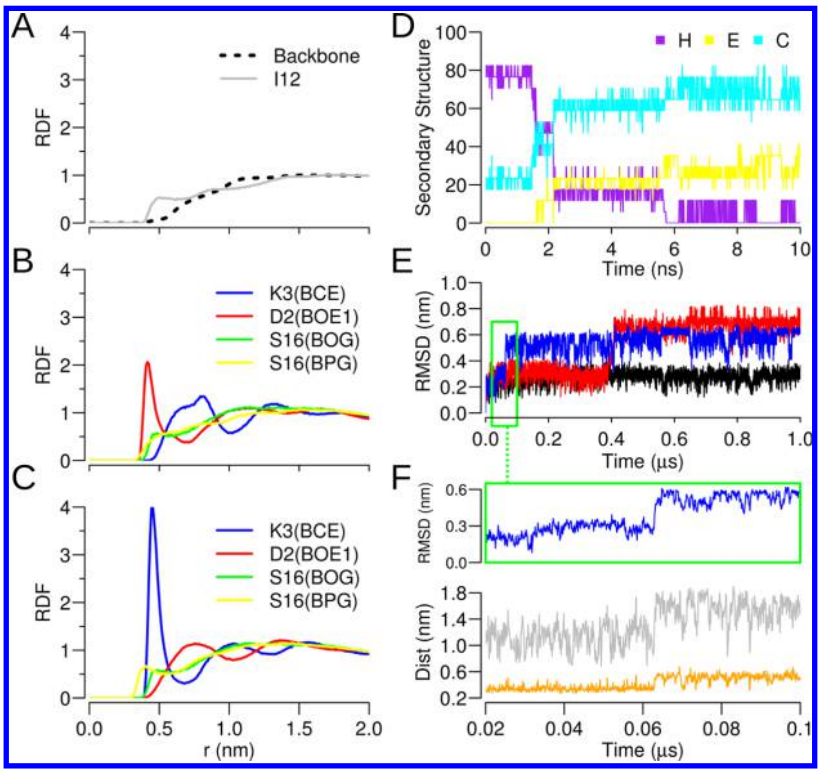


Figure 5. MD simulations of domain Ia of PLB. (A) RDF of WT4 around backbone and hydrophobic beads. The dashed line represents the distribution of WT4 beads around backbone beads corresponding to residues 2 to 16, while the continuous line is calculated for the side-chain bead of Ile12. (B) RDF of positive beads of WT4 around BOE1 in Asp2 (red), BCE in Lys3 (blue), BOG in Ser16 (green), and BPG in Ser16 (yellow). (C) Same as panel B for negative WT4 beads. (D) Secondary structure content of a 17-mer poly glycine started from a helical configuration. (E) RMSD calculated on systems at 300 K and 0.2 M NaCl (black), 330 K and 0.2 M NaCl (red), and 300 K with no added salts (blue, only 3 Cl− ions are added to ensure electroneutrality). (F) Top: Close up on the RMSD trajectory of PLB at low ionic strength in the moment of the conformational jump. Bottom: Distances between BCZ beads of Arg9 and 14 (gray) and backbone's GO and GN beads of Arg9 and 13, respectively (orange).

spontaneously within the ns time scale. These characteristic times are in good agreement with ns dynamics reported by single-molecule FRET and photoinduced electron transfer fluorescence experiments on glycine-rich peptides of similar length.[45]

(ii) As a second test, we attempt the thermal unfolding of domain Ia of PLB. To this goal, we use the starting configuration of the first simulation but coupled it to a reference temperature of 330 K. The dynamics of this system during the first 0.4 μs is comparable to that of the system at 300 K. However, after 0.4 μs of simulation, the helix undergoes a major break. As a consequence, the RMSD increases steeply undergoing large oscillations between 0.5 and 0.85 nm from the initial conformer (Figure 5E). This effect is accompanied by a marked reduction of the helical content. The average secondary structure content calculated during the last 0.6 μs is 40%, 13.5%, and 52.5% for helix, extended, and coil conformations, respectively.

(iii) A more challenging test for our force field is to assess its ability to capture ionic strength effects. We have previously shown that WT4 can adequately reproduce electrolytic properties as screening, osmotic pressure, Bjerrum and Debye lengths, etc., which result in correct concentration profiles, ion specificity, and local conformational changes observed in high-resolution X-ray structures of DNA.[26,32] The cytoplasmic domain of PLB seems a suitable test case as its stability has been shown to depend on the electrolytic content of the solution. Therefore, we set up a simulation of domain Ia of PLB at 300 K using the same starting conformer but in the absence of added salts. During the first 60 ns, the conformational sampling of the peptide is indistinguishable from its counterpart with RMSD peaks below 0.3 nm. After this period, the peptide helix experiences a major break, which is evident from the RMSD (Figure 5E). This produces a partially unfolded conformation with only a ∼40% of average helical content. Analysis of the trajectory indicates that the molecular cause of the helix rupture resides in the strong electrostatic repulsion between neighboring charged residues. In a helical context, the side chains of Arg9, 13, and 14 (Figure 4A) experience a strong reciprocal electrostatic repulsion. In the absence of proper electrostatic screening, this repulsion is hardly compatible with a helical structure. Because Arg13 and 14 are consecutive in sequence, the high electrostatic potential energy is relaxed by breaking the helix and separating the positive moieties. Indeed, the observed rise in RMSD is concomitant with the rupture of the H-bond interaction between backbone beads of Arg9 and 13, which leads to a separation of the side chains of Arg9 and 14 (Figure 5F). Conversely, the distances between the side chains of Arg9 and 14 during the simulation conducted at higher ionic strength are maintained at 0.9 nm (s.d. 0.2 nm). Taking into account that the Bjerrum length measured at 300 K in a solution of 0.2 M of NaCl in WT4 is of 0.57 nm,[26] it can be concluded that the electrostatic screening imposed by the

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

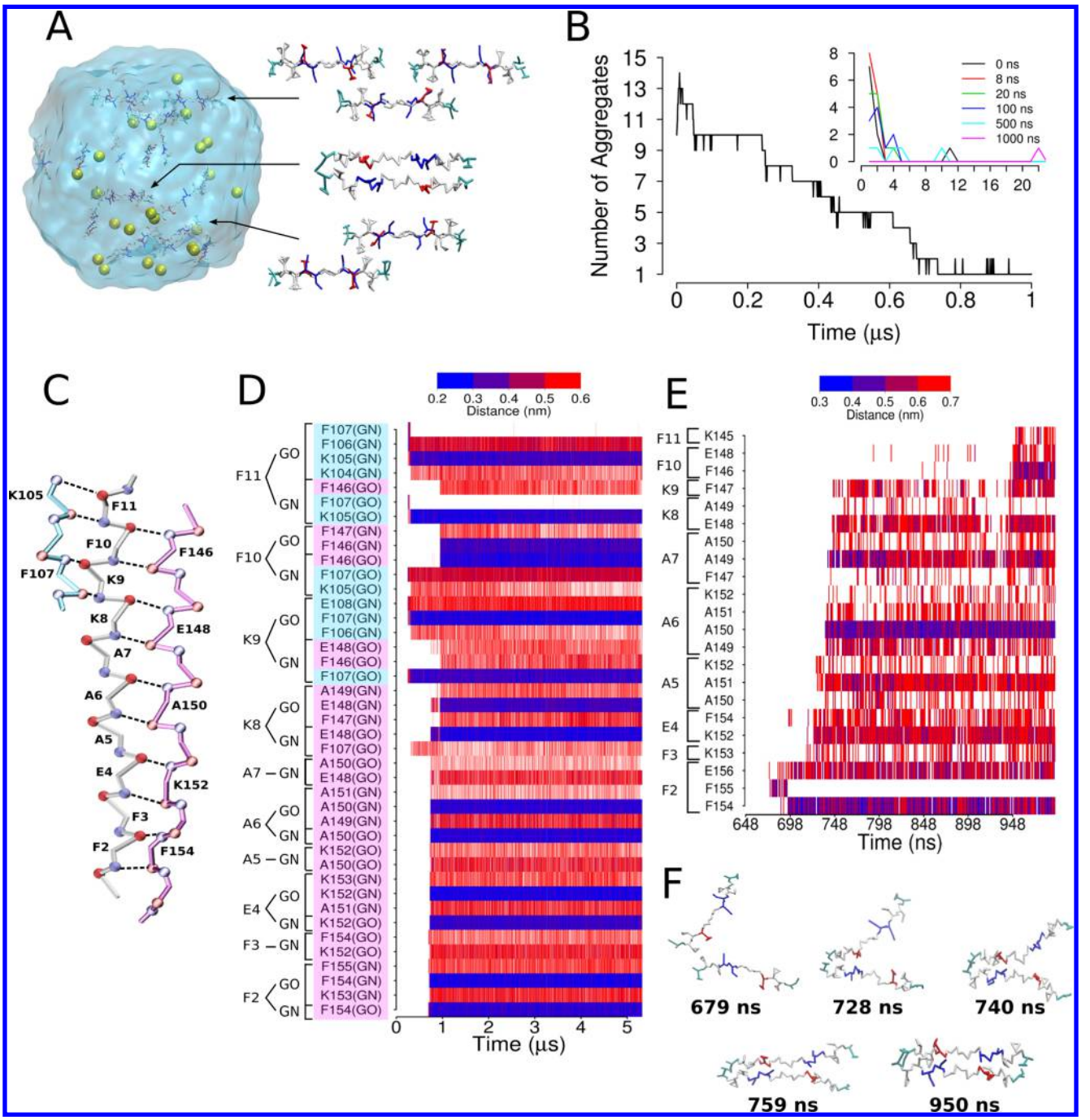


Figure 6. Self-aggregation of microglobulin β-amyloid peptides. (A) Simulation box of test case 2 showing the initial distribution of peptides colored by residue type as in Figure 4 with terminal residues in cyan. The position of the preassembled aggregates and counterions is shown. (B) Number of aggregates as a function of simulation time during the first microsecond. The inset shows the number of aggregates vs the number of components of each cluster at different simulation times. (C) Snapshot showing the aggregation of a triplet of peptides. The backbone of each peptide is shown in different colors. H-bonds are indicated with dashed lines. (D) H-bond interactions between the central and flanking peptides are identified by background shading with colors as in panel C. The instantaneous distances between GN and GO beads defining the β-sheet are shown on the right side. White colors correspond to distances higher that 0.6 nm. (E) Close up on the zippering process between peptides 1−12 and 144−156. Distances are measured between the center of mass of each residue. (F) Series of snapshots along the zippering of peptides 1−12 and 144−156 at different simulation times.

simple electrolytes present in the solution is sufficient to overcome the electrostatic repulsion, leading to structural stabilization of the cytoplasmic domain of PLB.

Taken all together, the results presented for this test case suggest that the SIRAH force field can qualitatively capture the main physicochemical aspects of aqueous solvation and interactions between functional groups in amino acids without any apparent conformational bias.

Test Case 2: Self-Aggregation of Microglobulin β-Amyloid Peptides. Amyloid fibrils are at the base of many lethal diseases, including Alzheimer's and spongiform encephalopathies (i.e., Creutzfeld−Jakob disease). In each disease, particular polypeptides form ordered, insoluble, and extended

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

fibrils that self-assemble in tissues, triggering different pathologies.[46−48] Simulating the aggregation processes that involves the formation of stable β-strand conformations and the unbiased self-assembly of β-sheets structures challenges the ability of our force field. For this task, we used the 12-mer amyloidgenic peptide KFFEAAAKKFFE, which has been studied by means of X-ray and electron diffraction.[49] In this peptide, the formation of ordered amyloid-like fibrils composed of antiparallel β-sheets is enhanced by $\pi-\pi$ stacking between adjacent phenylalanine rings and electrostatic interactions between charged residues (glutamic acid and lysine).

We set up a simulation box containing 10 isolated peptides plus three preassembled aggregates (one hexamer, one tetramer and one dimer), 4331 WT4 molecules, and 22 counterions (Figure 6A). The coordinates of the single peptides as well as those of tetramers and a hexamer are taken from the X-ray structure (PDB id: 2BFI).[49] Individual residues are numbered sequentially, so that peptide 1−12 contains residues from 1 to 12, etc.

The system shown in Figure 6A undergoes energy minimization, equilibration, and 5 μs of production MD. Positional restraints are not used at any point of the simulation protocol. Already at the equilibration, the preassembled aggregates dissociate, perhaps owing to the lack of stabilization provided by the crystal environment. Nevertheless, single peptides rapidly engage interactions with other partners. The initial configuration of the production phase consists of 10 aggregates that evolve rapidly, toward 12−14 during the first 50 ns, approximately (Figure 6B). Throughout the first microsecond, the total number of aggregates decreases almost monotonically defining intermediate conformations with occurrence times ranging from a few ns to over 0.2 μs. Small aggregates progressively fuse into higher size clusters, reaching saturation (one single aggregate of 22 peptides) after one microsecond of simulation.

To gain further insight into the self-assembly process, we monitored the formation of interstrand H-bonds interactions. For shortness, we focus on the aggregation process around peptide 1−12. The first contact occurs at nearly 0.3 μs between the backbone beads of residues K9 and F11 with F107 and K105, respectively (Figure 6C). After ∼0.7 μs, peptide 145−156 contacts peptide 1−12 initiating an antiparallel zippering to form a β-strand that remains stable for the rest of the simulation (Figure 6D). The alternation in distance profiles between backbone beads shown in Figure 6D is nicely coincident with that expected for an antiparallel β-sheet.

A more detailed analysis indicates that the first contact occurs at 679 ns between residues F2 and F155 (Figure 6E). Approximately 50 ns after such contact, a salt bridge forms between E4 and K152. Rapidly after that interbackbone H-bonds form between the three central alanines of each chain (740 ns). This is followed by the formation of a second salt bridge between K8 and E148 (759 ns). Finally, close to the end of the first microsecond, F10 pairs with F146 forming a second hydrophobic contact that "zips up" the β-strand. These interactions are maintained until the end of the simulation. As expected from solvent competition, salt bridges show more noticeable fluctuations than hydrophobic contacts (Figure 6E).

The stabilization provided by hydrophobic contacts between phenylalanine rings and salt bridge interactions between lysine and glutamate residues is consistent with previously reported data,[49] sustaining the ability of the model not only to describe the formation of backbone H-bonds that lead to β-sheets configurations but also to feature selective interaction patterns according to the physicochemical properties of different residues.

Test Case 3:Alpha/Beta Protein TOP7. As a first test for the structural stability of a protein, we select the single domain alpha/beta protein Top7 (PDB id: 1QYS). This protein was originally designed using computational techniques and was showed to have an extremely stable folding.[50] Indeed, its structural stability has been recently proven by inserting different HIV-1 epitopes into the Top7 scaffold without perturbing the global folding.[51] This protein contains the most common and stable secondary structure motifs, i.e., two α-helices and five β-strands arranged in an antiparallel β-sheet (Figure 7A). After protonation and mapping of FG coordinates to CG, we set up a system containing the protein solvated with 1033 WT4 molecules and 30 NaCl ionic couples representing a rough ionic strength of ∼0.15 M. Because the first and last residues in the sequence are present in the crystal but not

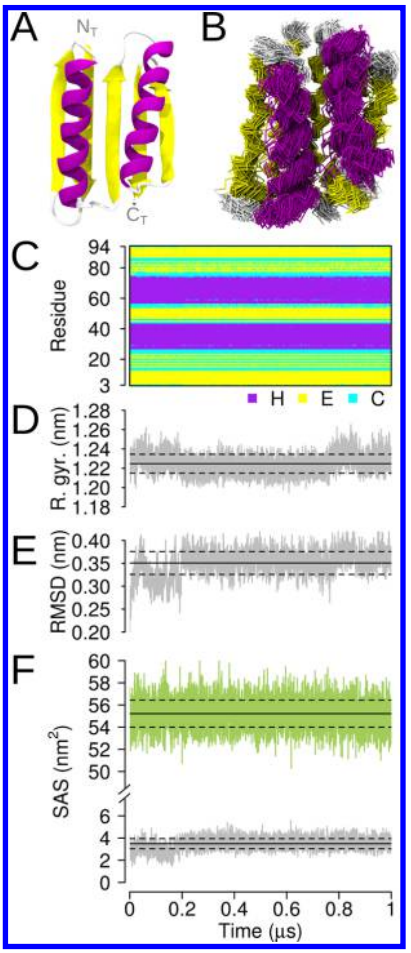


Figure 7. MD simulations of Top7 protein. (A) Cartoon representation of the FG structure of Top7 showing the secondary structure elements (helix, purple; extended, yellow; coil, white). Amino and carboxyl termini (NT and CT) are indicated. (B) Backbone superposition of 100 snapshots taken every 10 ns during the trajectory colored as in panel A. (C) Secondary structure per residue as a function of time. (D and E) Radius of gyration and RMSD as a function of the simulation time. (F) Solvent accessible surface (SAS) of hydrophilic (top) and hydrophobic (bottom) residues. In panels D−F, black solid and dashed lines indicate the average and standard deviations values, respectively.

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

solved in the experimental electronic density, the capping amino acids are considered as neutral (not zwitterionic). Using this set up, we carry out a production run of 1 μs. Visual inspection of the superposition of 100 frames equally separated along the simulation time suggests a good conservation of the secondary and tertiary structure of the protein upon CG dynamics, specially taking into account the absence of topological constraints in the force field (Figure 7B). In fact, it is immediate to recognize the helical and extended segments as a function of the sequence and simulation time (Figure 7C). As a result of thermal motion, the structure undergoes small fluctuations, which are more pronounced near the extremities of structural elements. In particular, the strands spanning residues 15−24 and 77−84 show a more labile character, which is originated by the topology of the protein. Inspection of Figure 7A indicates that these two segments are indeed at the flanks of the β-sheet and hence less stabilized by the protein scaffold. Quantitatively, the secondary structure calculation on the initial coordinates results in H: 38.0%, E: 43.5% and C: 18.5%. These values experience small variations over the dynamics with averages of H, 37.4% (s.d. 0.8); E, 39.2% (s.d. 2.2); and C, 23.4% (s.d. 2.3).

To acquire a more comprehensive assessment of the quality of SIRAH, we adopted a series of descriptors to characterize the structure and dynamics simulated for this and other systems.

As a first indicator of the global molecular shape, we calculate the radius of gyration, which after minor variations within the first 0.2 μs adopts a value very close to the original (1.21 nm) oscillating around 1.22 nm. (Figure 7D). In line with these results, the RMSD suggests that the overall folding of the protein is well maintained, although the RMSD experiences a steep increase during the first 50 ns, followed by smaller fluctuations during the next 0.15 μs to then reach a plateau with an average value of 0.36 nm (s.d. 0.02) (Figure 7E).

Despite the good behavior of these descriptors, it is still possible that the protein remains in a compact state with Cα relatively close to the initial conformation but with flipped amino acids side chains exposing hydrophobic parts to the solvent. To exclude this possibility, we calculated the SAS of hydrophilic and hydrophobic residues. This is an important control as significant variations in SAS would be indicative of a poorly reproduced hydrophobic/hydrophilic balance in our CG water model. As shown in Figure 7F, hydrophilic and hydrophobic surfaces are well maintained with a modest increase in the hydrophobic SAS, which stabilizes after ~0.2 μs. Average values calculated before and after that point differ 0.8 nm^2 from the initial value. This represents a very small difference considering that the hydrophobic surface of an atomistic alanine is 1.1 nm^2, suggesting a good reproduction of the hydrophobic effect in WT4.

Finally, we sought to monitor the conservation of native contacts along the dynamics as a sensible indicator of the faithful conservation of the protein topology. If the analysis is performed skipping the first five neighboring residues to avoid contacts directly related with the secondary structure (non-contiguous contacts), we obtain a conservation of 75% (s.d. 3) with an accuracy of 64% (s.d. 3). However, considering all possible contacts (global contacts), the conservation rises to 89% (s.d. 1) with an accuracy of 84% (s.d. 1).

All together, the conformational and dynamical descriptors analyzed indicate good behavior of the force field for this very stable and well-folded protein.

Test Case 4: Cyclic Nucleotide Binding Domain of HCN Channel. The cyclic nucleotide binding domain (CNBD) is a conserved protein module of nearly 120 amino acids, comprising helical elements and an eight-stranded β-barrel, which serves as a binding site for cyclic nucleotides. A conserved arginine is present within the binding site to establish a salt bridge interaction with the exocyclic phosphate of the cyclic nucleotide upon binding.[52] This arginine is partially buried but accessible to the solvent in the apo state (Figure 8A). Structural determinations of these protein domains in absence of cAMP remained elusive to experimental methods for many years owing to its intrinsic flexibility. Recently, the structure of the CNBD of the HCN channel has been solved by a combination of spectroscopic techniques (PDB id: 2MNG).[53] This structure was chosen as a test case because of the challenging topology of the β-barrel and to ascertain the capability of the force field to reproduce a flexible fold and the local solvation of a partially buried Arginine.

In close similarity with the previous case, the secondary and tertiary structure of this protein is well conserved, as shown from the comparison between Figure 8A and B. The variation in the radius of gyration remains below 5% of its initial value, and the conservation of global contacts scores 84% (s.d. 3) with an accuracy of 88% (s.d. 1). Calculation of the RMSD indicates that the structure stabilizes within the first 0.1 μs (Figure 8C). Along the trajectory, the protein explores conformations that are almost 0.6 nm far from the initial state (model 1 in the NMR family). However, to establish a proper contrast with experimental data, we compare the structural plasticity of the protein along the MD trajectory with all the conformations in the NMR ensemble.[53] As shown in Figure 8C top, the protein visits a series of states that are alternatively closer to one or more NMR conformers. After 0.3 μs, the CNBD separates from the initial conformer by more than 0.5 nm but samples conformations closer to NMR models 5 and 7. Calculating the minimum RMSD to any of the NMR conformers indicates that the protein remains at nearly 0.4 nm from at least one of the experimental structures (Figure 8C middle). This puts forward the capability of SIRAH to sample different energetic states on a potential energy surface, roughly reproducing the experimental ensemble of conformations.

Analysis of the secondary structure during the simulation time suggests a good global conservation with some marked local fluctuations (Figure 8C, bottom). In particular, we notice that the region spanning residues 640−660 shows poorer secondary structure conservation. However, it is interesting to notice that this stretch contains the β4−5 loop, which is the least conserved region in the entire protein family.[54] Moreover, NMR studies report a local minimum in the HN NOE profile for residues 645−655,[53] and atomistic simulations have shown high fluctuations for the same residues.[55,56]

Finally, we sought to investigate the solvation of R669, which is conserved in this protein module and is responsible for the affinity to cyclic nucleotides. Given the relatively large dimensions of WT4, it is not obvious that this arginine partially buried in a hydrophobic environment can be solvated without perturbing the protein scaffold. However, monitoring the arrangement of WT4 molecules around R669, we observe that WT4 can penetrate into the cavity and interact with this arginine (Figure 8D). This interaction is clearly specific and driven by electrostatics as negative beads of WT4 present a preferential orientation toward the side chain of R669. This reinforces the role of the electrostatics in the proper orientation

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

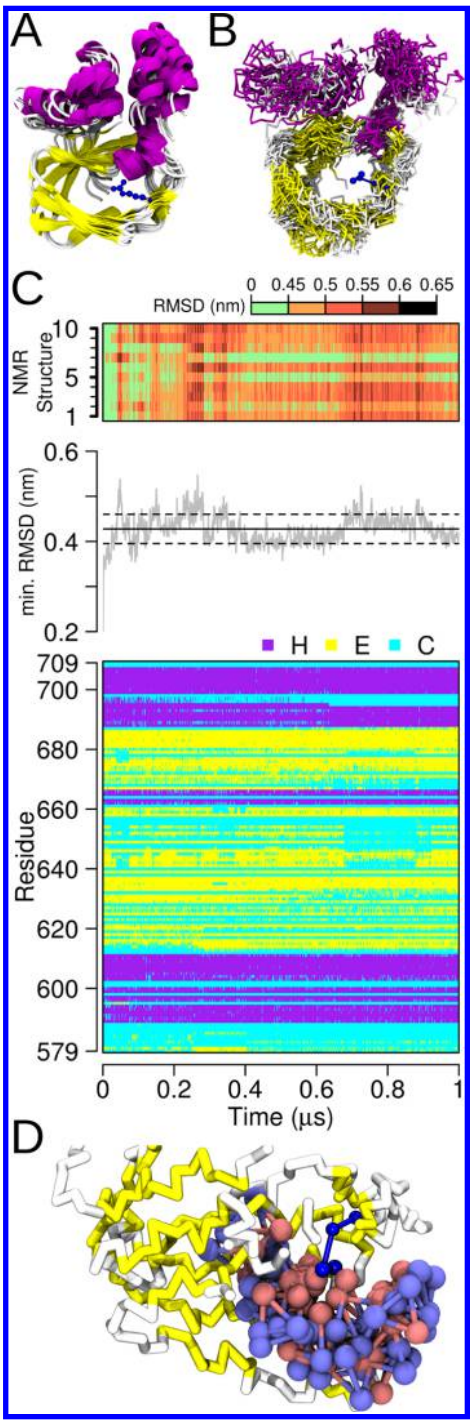


Figure 8. MD simulations of HCN's CNBD. (A) Superposition of all conformers in the NMR structure (PDB id: 2MNG) colored by secondary structure. R669 within the β-barrel binding site is shown in blue balls and sticks. (B) Backbone superposition of 100 snapshots taken every 10 ns during the trajectory. R669 is represented as in panel A. (C) Top: Color matrix of RMSD values along the CG trajectory against the NMR ensemble. Middle: Minimum RMSD from any of the NMR conformers. Continuous and dashed lines indicate the average and standard deviations. Bottom: Secondary structure assignment. (D) Superposition of WT4 molecules around 0.5 nm of the BCZ bead of R669 colored by charge (positive, blue; negative, red). The β-barrel and R669 are shown in the initial state, while WT4 snapshots are taken every 100 ns.

of the solvent observed for charged residues shown in Figure 5B and C for test case 1.

Test Case 5: HP1 Chromo-Shadow Dimer in Complex with CAF1 Peptide. As a more challenging test case, we undertake the study of a trimeric complex as that constituted by the chromo-shadow (CS) domain of the heterochromatin protein 1 (HP1) bound to an extended peptide from the chromatin assembly factor-1 (CAF1) determined by NMR spectroscopy.[57] The CS domain is present at the C-terminal of the HP1 and is named after its high homology to the chromo (CHROmatin MOdifier) domain.[58] Upon dimerization, the CS homodimer creates a binding site that recognizes small peptides carrying the consensus motif PXVXL. These peptides bind with high affinity in an extended conformation to a cleft formed by C-terminal residues of the CS dimer forming a trimeric β-sheet (Figure 9A). Each component of this trimeric complex contains a rigid core with highly flexible and solvated terminals.

This domain dimerizes through a nontrivial interface that involves hydrophobic, aromatic, polar, and charged residues in α-helix and β-sheet conformations. Hence, the correct representation of the intramolecular interactions shown in the previous test cases may not be enough to properly capture the noncovalently bonded forces holding the trimeric complex. Moreover, this requires a fine specificity in the formation of interchain H-bonds and hydrophobic protein−protein interfaces in the correct relative position.

The CG simulation shows a stable trajectory with an average radius of gyration of the dimer of 1.8 nm, which compares very well with the 1.9 nm calculated on the NMR set. The average RMSD of the dimer calculated against the NMR ensemble results in 0.47 nm (s.d. 0.02). Analogously, we found a comparable structural variability for each CS monomer with RMSD values of 0.45 nm (s.d. 0.02) and 0.40 nm (s.d. 0.03). Indeed, the global shape is well preserved as pointed out by snapshots taken along the trajectory (Figure 9A). Consistently, the secondary structure calculated along the trajectory is conserved in all three components of the complex. Both monomers present the same global pattern of secondary structure along the simulation with unstructured N- and C-tails, a β-sheet, and two short helical elements, which mediate the dimerization interface. Additionally, the binding peptide remains in an extended β-strand strand conformation (Figure 9B).

Analysis of global contacts at the interface residues reveals a conservation of 55% (s.d. 4) with an accuracy of 26% (s.d. 2.6), suggesting that lumping the side chain of hydrophobic amino acids into one single bulkier CG bead (Figure 2) may weaken the specificity of some interactions. To further assess this, we calculate the protein−protein interface. The NMR family is characterized by an average total protein−protein interface of 10.7 nm^2 (s.d. 0.5) in comparison with an average of 15.2 (s.d. 0.8) obtained along the CG simulation. This effect is, however, limited to the dimerization interface, as calculation of global contacts lead to 80% of conservation with an accuracy of 76%. Hence, although fine details of the protein−protein interface can be missed, the global topology is well conserved. In agreement, the average total SAS of the complex for the NMR family is 115.1 nm^2 (s.d. 1.9), while its counterpart in the simulation is 112.1 nm^2 (s.d. 2.2).

Regarding the interaction of the CS dimer with the CAF1 peptide, the by-residue SAS and the secondary structure analysis of the CAF1 peptide shows that the PXVXL region remains buried in the binding site cavity and interacting in an

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

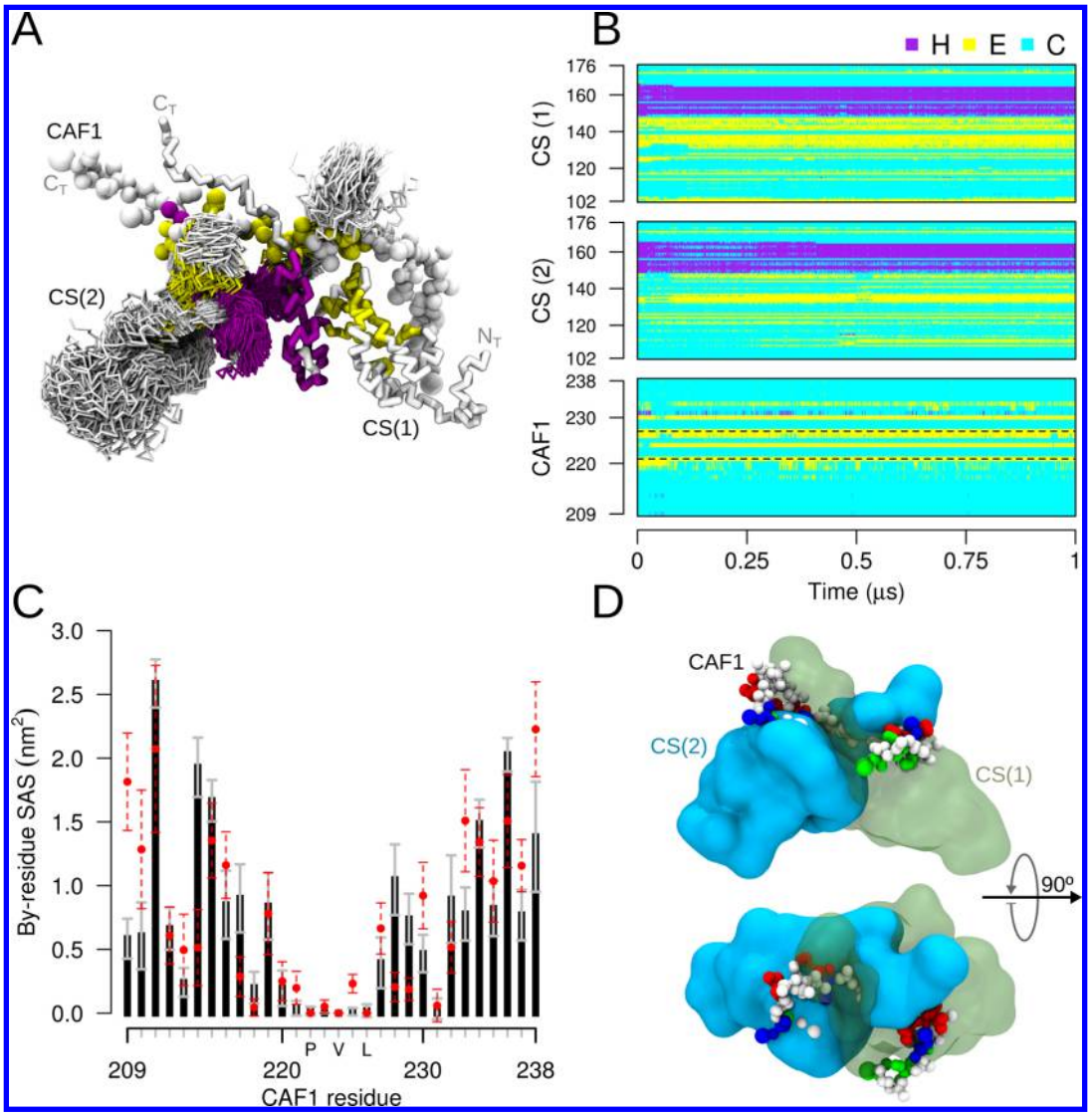


Figure 9. HP1 CS dimer in complex with CAF1 peptide. (A) The backbone's initial conformation of monomer CS(1) is represented by thick tubes, while an ensemble of conformations for monomer CS(2), taken one each 10 ns along the of trajectory, is represented by thin tubes. The initial conformation of the CAF1 peptide is shown as balls and sticks. All residues are colored according to the secondary structure. Amino and carboxy terminal residues (NT and CT) are indicated. (B) Variation of secondary structure along 1 μs of trajectory. The location of the PXVXL motif within the CAF1 sequence is indicated by dashed lines. (C) By-residue SAS calculation on the CAF1 peptide. Columns and solid line error bars correspond to average and standard deviation values during the CG simulation. Circles and dashed error bars indicate the analogous quantity measured from the NMR family of conformers. Residues belonging to the signature motif (PXVXL) are indicated. (D) Molecular representation of the final coordinate from the CG trajectory. The CS monomers are represented as cyan and green surfaces, while the CAF1 peptide is shown in balls and sticks, and it is colored by residue type (basic, blue; acidic, red; polar, green; and nonpolar, white).

extended beta conformation, in good agreement with the experimental data[57] (Figure 9B and C, respectively). Moreover, we retrieve a quasi-quantitative agreement between the exposed surfaces of the portion of the CAF1 peptide outside the signature motif. This region is mostly unstructured and exposed to the solvent with the C-terminal tails of the CS monomers embracing the CAF1 peptide (Figure 9D) as it is needed for these regions to actively participate in the ligand recognition.[59]

Test Case 6: The SNARE Complex. The SNARE proteins are the main constituents of the synaptic vesicle fusion machinery. The cytoplasmic domains of the three SNARE proteins, VAMP (vesicle-associated membrane protein, also known as synaptobrevin), syntaxin, and SNAP-25, spontaneously form a parallel coiled-coil four-helix bundle, providing sufficient free energy to drive membrane fusion.[60−62] Continuous helical segments of about 60 residues (one from VAMP, one from syntaxin, and two from SNAP-25) form the fully zippered SNARE complex (Figure 10A).

This system is chosen as a final test case because if small systematic errors are present in our force field they should accumulate and become evident in the very long helices of the SNARE complex, impairing the stability of the heterotetramer upon dynamics. This system is constructed from the PDB structure 1KIL[63] removing the chain E corresponding to complexin. This simulation results in a very stable trajectory with a RMSD that stabilized already after the first 10 ns oscillating around values of ∼0.3 nm. Along the simulation, the secondary structure is very well maintained, and the whole stability of the 4-helix bundle is not compromised as shown by the superposition of snapshots in Figure 10B.

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

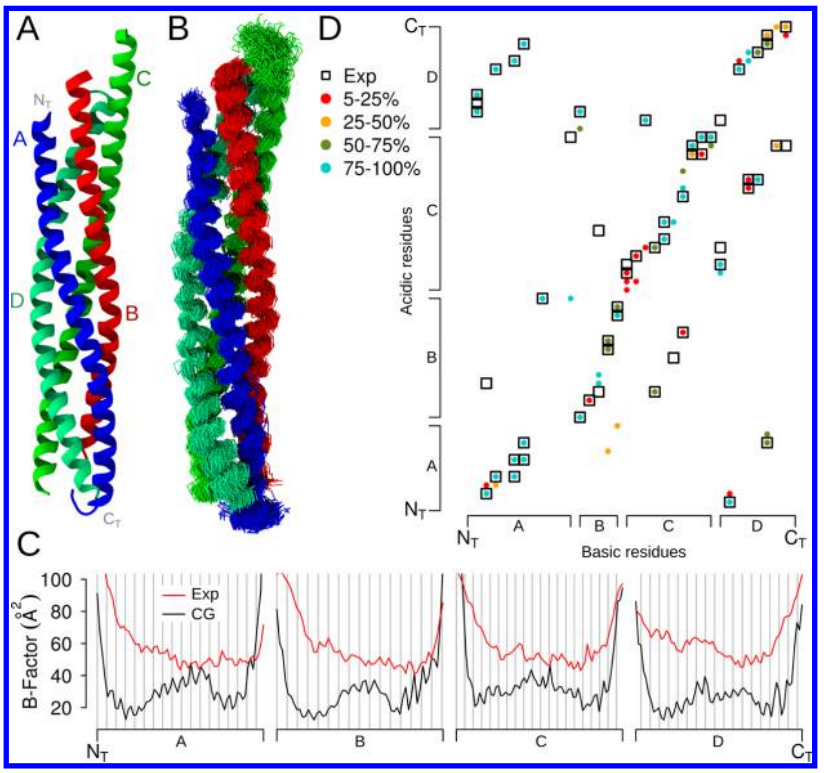


Figure 10. MD simulations of the SNARE complex. (A) Cartoon representation of the SNARE complex colored by chain (A, VAMP; B, syntaxin; C, SNAP-25 N-terminal helix, and D, SNAP-25 C-terminal helix). (B) Superposition of 100 snapshots taken every 10 ns during the trajectory colored as in panel A. (C) B-factors on the Cα beads calculated along the dynamics plotted against residue number for the four helices in the complex. The vertical gray lines are drawn every 3.6 residues to indicate the helical periodicity (see text). Experimental values are shown in red. (D) Global pattern of salt bridges in the SNARE complex. The vertical axis indicates the position of acidic residues (aspartic and glutamic acid) sequentially ordered versus basic residues (arginine and lysine). Dots represent the existence of salt bridges at any point of the simulation and are color coded by the percentage of occurrence along the trajectory. The experimentally observed salt bridges are presented as empty squares.

Calculation of the B-factors on the Cα beads measured along the simulation indicates that as expected termini are more mobile that the central portion of the protein (Figure 10C). Moreover, each chain presents a fine structure with a serrated shape. A deeper consideration of this pattern indicates that the alternation of values nicely coincides with the 3.6 periodicity of a α-helix (gray lines, Figure 10C) and is imposed by the quaternary structure. Exposed residues show higher B-factors, while buried ones do the opposite. Although a direct comparison with the crystal structure would not be strictly correct because of packing effects, temperature, buffer conditions, etc., it is still possible to find a rough correlation with the experimental B-factors (Figure 10C).

The conservation of native contacts along the simulation considering noncontiguous interactions is 67% (s.d. 3), while the global contacts results in 92% (s.d. 1). Considering the predominantly helical topology of the complex, noncontiguous contacts are essentially related to the protein−protein interface. Hence, the difference between global and noncontiguous conservation is in line with the conclusion drawn in the previous test case. There, we observe a certain degree of promiscuity within the protein−protein interface. However, the protein−protein interfaces between each helix and the rest of the complex calculated along the simulation, differ at most in a 5% from the experimental values.

Finally, we turn our attention to the electrostatic interactions in the complex. To this aim, we calculated the salt bridges existing in the crystallographic structure and compare them with those conserved and/or arising during the simulation. A global view of the salt bridge network on the SNARE complex is presented in Figure 10D. A comparison indicates that 92% of the salt bridges measured from the experimental data are conserved during the simulation with different occurrence times. We also observe the transient formation of additional interactions, mainly concentrated on helix C (SNAP-25 N-terminal). These interactions have in general low occurrence times and are concentrated in the diagonal region of the graph, being indicative of intramolecular bridges. Off-diagonal elements are conserved, supporting the good reproduction of the quaternary structure.

Testing the Robustness of SIRAH. Finally, aimed to acquire a more general vision of the performance of SIRAH, we carried out a set of simulations of different proteins. These structures are selected from the PDB with the following criteria: (i) Structures contain only a protein without any other small ligand or macromolecular species. (ii) There are no disulfide bonds or modified amino acids. (iii) Molecules are monomeric with one single chain per asymmetric unit. This minimally representative group of systems is protonated at pH 7 using the pdb 2pqr server (http://nbcr-222.ucsd.edu/pdb2pqr_1.8/). Without any further checking, proteins in their zwitterionic form are solvated in an octahedral box adding 1.5 nm of WT4 in each direction and simulated at 300 K and 1 atm in the presence of 0.15 M of NaCl for 1 μs. For the sake of shortness, results are summarized in Table 3. In all the cases, the quantities are calculated on the experimental structure and compared against the average of the last 0.1 μs of each trajectory.

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 3. Robustness of SIRAH on Different Proteins[a]

PDB	N	RMSD (nm)	radius of gyration (nm)	SAS (nm^2) Hphob. Hphil.	secondary structure H, E, C	contacts conservation/ accuracy
1TQG	105	0.38(0.04)	[1.38] 1.41(0.01)	[6.4] 80.3(1.3); [59.7] 54.0(1.1)	[86.7] 73.8(1.1); [1.9] 3.1(1.6); [11.4] 23.1(2.2)	86.8(0.6)/ 80.3(1.3)
1R69	63	0.87(0.10)	[1.01] 1.34(0.03)	[3.2]10.3(0.7); [38.2]42.1(1.2)	[69.8]58.4(2.4); [0.0]2.3(1.6); [30.2]39.4(3.1)	83.8(1.0)/ 87.9(1.4)
2CKX	83	0.71(0.07)	[1.17] 1.31(0.01)	[8.2] 9.9(0.5); [46.5] 49.6(1.1)	[66.3] 58.2(2.0); [6.0] 10.7(1.7); [27.7] 31.0(2.7)	79.1(0.5)/ 72.9(1.2)
1BKR	108	0.60(0.08)	[1.24] 1.31(0.03)	[6.1] 11.4(1.1); [57.4] 56.2(1.5)	[58.3] 55.2(1.5); [3.7] 6.4(1.1); [38.0] 38.4(1.8)	77.2(1.0)/ 73.7(1.4)
1RA4	117	0.36(0.04)	[1.23] 1.24(0.01)	[6.2] 8.2(0.6); [59.7] 63.1(1.5)	[48.7] 45.6(1.3); [19.7] 15.1(1.3); [31.6] 39.3(1.8)	81.3(0.9)/ 76.4(0.9)
2KYR	108	0.58(0.06)	[1.27] 1.38(0.01)	[8.2] 13.4(0.6); [56.9] 64.2(1.5)	[46.3] 44.2(1.2); [32.4] 21.7(2.0); [21.3] 34.1(2.4)	74.8(1.0)/ 81.3(1.2)
2VIM	104	0.55(0.06)	[1.22] 1.35(0.02)	[5.6] 11.4(1.0); [52.7] 57.4(1.5)	[46.2] 40.6(0.8;) [36.5] 29.6(2.2); [17.3] 29.8(2.3)	80.4(1.3)/ 78.2(1.8)
1CRN	46	0.76(0.08)	[0.97] 1.13(0.04)	[8.1] 13.1(0.8); [24.1] 28.7(1.0)	[45.7] 40.2(1.5;) [23.9] 10.9(3.4); [30.4] 48.8(3.5)	80.4(1.4)/ 79.4(2.3)
1ORC	64	0.49(0.05)	[1.08] 1.10(0.02)	[4.9] 7.1(0.6); [43.3] 43.2(1.1)	[40.6] 33.4(1.6); [35.9] 21.5(2.7); [23.4] 45.1(3.1)	79.1(1.5)/ 73.1(2.4)
1PGB	56	0.41(0.06)	[1.03] 1.06(0.01)	[2.9] 3.2(0.4); [37.9] 41.6(1.1)	[26.8] 24.8(1.3); [51.8] 37.5(4.6); [21.4] 37.7(4.8)	88.5(1.9)/ 80.7(2.3)
2IUG	110	0.55(0.06)	[1.28] 1.39(0.02)	[8.0] 10.2(0.7); [57.1] 63.6(1.6)	[22.7] 21.1(1.1); [37.3] 18.6(2.7); [40.0] 60.4(3.0)	74.0(1.3)/ 77.3(1.2)
1QYO (E222N)	236	0.50(0.05)	[1.75] 1.80(0.01)	[8.2] 12.5(0.7); [105.1] 101.8(1.9)	[7.2] 8.7(0.6); [55.5] 55.0(1.7); [37.3] 36.3(1.8)	80.1(0.5)/ 77.3(0.6)
1OPS	64	0.56(0.06)	[1.01] 1.05(0.01)	[8.9] 10.2(0.5); [30.4] 35.4(1.1)	[4.7] 0.5(0.7); [56.2] 34.1(3.8); [39.1] 65.5(3.8)	69.6(1.1)/ 71.0(1.8)
1GYV	120	0.35(0.04)	[1.44] 1.50(0.01)	[9.1] 11.5(0.7); [60.1] 62.6(1.1)	[1.7] 1.0(0.6); [61.7] 52.8(2.7); [36.7] 46.2(2.8)	82.6(1.0)/ 81.0(0.9)
2E3H	76	0.55(0.06)	[1.12] 1.16(0.03)	[8.0] 9.7(0.8); [43.3] 44.1(1.7)	[0.0] 1.0(0.6); [48.7] 45.8(5.1); [51.3] 53.2(5.1)	73.7(1.3)/ 79.5(2.3)
1PWT	61	0.60(0.07)	[1.03] 1.15(0.02)	[5.1] 9.3(0.6); [43.5] 40.0(1.0)	[0.0] 0.6(0.8); [60.7] 34.9(3.2); [39.3] 64.6(3.4)	68.2(1.4)/ 81.1(1.4)
2O31	67	0.55(0.07)	[1.05] 1.26(0.03)	[9.5] 13.5(0.7); [40.1] 42.1(1.3)	[0.0] 0.2(0.5); [62.7] 36.3(4.2); [37.3] 63.5(4.3)	73.5(2.3)/ 83.5(1.5)
1XX8 (150 mM NaCl)	66	0.63(0.09)	[1.12] 1.14(0.03)	[3.7] 5.0(0.8); [49.2] 50.6(1.0)	[16.7] 10.7(1.7); [42.4] 34.8(3.9); [40.9] 54.5(4.3)	82.9(1.1)/ 81.7(2.0)
1XX8 (300 mM NaCl)	66	0.37(0.04)	[1.12] 1.06(0.01)	[3.7] 3.1(0.4); [49.2] 47.0(1.1)	[16.7] 13.6(1.1); [42.4] 42.0(3.1); [40.9] 44.3(3.3)	88.3(1.3)/ 81.3(1.7)

[a]N indicates the number of residues. Values reported correspond to the average calculated over the last 100 ns of each μs. Parentheses and square brackets indicate standard deviations and values calculated from the experimental structure, respectively.

In terms of RMSD, we observe that deviations from the initial structures range from nearly atomistic (0.35 nm for 1GYV) up to as high as 0.87 nm for 1R69 (see below). Nevertheless, a comprehensive comparison of the radius of gyration in all the cases indicates that the compaction of the proteins experiences less relevant variations. This is supported by relatively high conservation of global contacts for which the lowest value found is 68% for 1PWT. The low variations in the SAS values indicate a conservation of the hydrophobic core of the molecules with an acceptable balance between the hydrophobicity/hydrophilicity of amino acids with different physicochemical characteristics. Secondary structure elements show a variability that can be comparable with that expected from FG simulations, showing no apparent bias for any structural motif.

This set of simulations provides an estimation of the robustness of the force field because none of the systems was set up with optimal conditions such as temperature, ionic strength, protonation states, charges in the termini, etc. Similarly, gentler minimization/initialization protocols could significantly improve the outcome of the simulations. As a concrete example, we consider the simulation of the structure 1XX8 (last rows in Table 3). This structure corresponds to a DNA binding protein and has a very high charge density with a net charge of +6 in 66 residues. Simply doubling the ionic strength in the solvation box significantly improves the stability in all the molecular descriptors (compare two last rows in Table 3).

It is also important to recognize that there are still situations in which the performance of our force field is not satisfactory. For instance, the protein 1R69 is the worst case reported in Table 3. This protein comprises five short α-helices, two of these forming a helix-turn-helix motif. Analysis of the trajectory suggests that one of the causes of the partial unfolding is the solvation of a partially buried salt bridge between R10 and E35. Owing to the high granularity of WT4, it cannot occupy single-water cavities, failing to provide stabilizing interactions within small hydrophobic cavities. Additionally, the helix-turn-helix motif may present challenges to relax the tension in the reduced number of torsional degrees of freedom in CG models. These constitute clear limitations intrinsic of the CG approach and are difficult to capture without the use of specific constraints and have to be evaluated in each particular case. An example of possible solutions for this kind of problems is the simulation of the protein 1QYO, which is a variant of the green fluorescent protein that is not able to create a

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

chromophore. This protein contains nearly 30 single water molecules within the β-barrel and E222 pointing into the β-barrel. The impossibility of introducing WT4 molecules within the barrel results also in a badly unfolded protein (not shown). However, introducing the E222Q mutation to mimic a protonated glutamic acid, as suggested by PROPKA calculations,[64] produces a stable trajectory with well conserved overall molecular descriptors (Table 3).

CONCLUSIONS

In this paper, we present an extension of the SIRAH force field for unbiased MD simulations of peptides and proteins at the CG level using explicit solvation and long-range electrostatics. In line with our previous developments, all the interactions of this residue-based CG model are represented within a classical Hamiltonian, which is common to most MD simulation packages. Comparative simulations with fully atomistic systems containing an equivalent number or particles as in the test cases presented indicate a speed up of 2 orders of magnitude. The FG-to-CG mapping using positions of real atoms to place CG beads allows for the nearly atomistic identification of functional groups establishing interactions as H-bonds and salt bridges. This mapping strategy will facilitate the future development of modified amino acids to study the influence of phosphorylation, methylation, different protonation states, etc. The use of long-range electrostatics, permittivity of the solvent, and explicit presence of ions in the solution allows for including ionic strength effects that are very difficult to capture for CG approaches.

Clearly, this contribution paves the way for the study of protein−DNA complexes using SIRAH. Although preliminary simulations show encouraging results, statistic analysis against protein−DNA interfaces reported in the PDB reveals low discrimination in the interaction with the positive side chains. This would call for an optimization of the interaction between phosphate groups and the side chains of arginines and lysines. Work is underway in our group to address this problem and will be the subject of a forthcoming publication.

To facilitate the implementation of the SIRAH force field, a tarball containing a complete set of interaction topology files in Gromacs format (version 4.5.5) is freely available at http://www.sirahff.com. The documentation contained therein includes step-by-step tutorials, prestabilized WT4 solvation boxes, and mapping scripts to convert from PDB to CG and from CG to PSF formats, facilitating the visualization and analysis of molecular systems.

AUTHOR INFORMATION

Corresponding Author

*Tel/Fax: +598-2522 0910. E-mail: spantano@pasteur.edu.uy.

Author Contributions

L.D. and M.R.M. equally contributed to this work

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partially funded by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11. A.F.B. is beneficiary of a National Fellowship provided by ANII. M.R.M. and S.P. are researchers from the National Scientific Program of ANII (SNI). We thank Paolo Carloni and Anna Tramontano for very useful discussions.

REFERENCES

(1) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321*, 798−800.
(2) Saiz, L.; Klein, M. L. Computer simulation studies of model biological membranes. *Acc. Chem. Res.* **2002**, *35*, 482−489.
(3) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1 ed.; Taylor & Francis Group: New York, 2009; pp 1−455.
(4) Sansom, M. S.; Scott, K. A.; Bond, P. J. Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem. Soc. Trans.* **2008**, *36*, 27−32.
(5) Orsi, M.; Essex, J. W. The ELBA force field for coarse-grain modeling of lipid membranes. *PLoS One.* **2011**, *6*, e28637.
(6) Trylska, J.; Tozzini, V.; Chang, C. E.; McCammon, J. A. HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophys. J.* **2007**, *92*, 4179−4187.
(7) Sieradzan, A. K.; Niadzvedtski, A.; Scheraga, H. A.; Liwo, A. Revised backbone-virtual-bond-angle potentials to treat the L- and D-amino acid residues in the coarse-grained united residue (UNRES) force field. *J. Chem. Theory Comput.* **2014**, *10*, 2194−2203.
(8) Basdevant, N.; Borgis, D.; Ha-Duong, T. Modeling protein−protein recognition in solution using the coarse-grained force field SCORPION. *J. Chem. Theory Comput.* **2013**, *9*, 803−813.
(9) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(10) Pasi, M.; Lavery, R.; Ceres, N. PaLaCe: A coarse-grain protein model for studying mechanical properties. *J. Chem. Theory Comput.* **2013**, *9*, 785−793.
(11) Han, W.; Schulten, K. Further optimization of a hybrid united-atom and coarse-grained force field for folding simulations: Improved backbone hydration and interactions between charged side chains. *J. Chem. Theory Comput.* **2012**, *8*, 4413−4424.
(12) Arkhipov, A.; Yin, Y.; Schulten, K. Four-scale description of membrane sculpting by BAR domains. *Biophys. J.* **2008**, *95*, 2806−2821.
(13) Spiga, E.; Alemani, D.; Degiacomini, M. T.; Cascella, M.; Dal Peraro, M. Electrostatic-consistent coarse-grained potentials for molecular simulations of proteins. *J. Chem. Theory Comput.* **2013**, *9*, 3515−3526.
(14) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.* **2005**, *95*, 218102.
(15) López, C. A.; Rzepiela, A. J.; de Vries, A. H.; Dijkhuizen, P. H.; Hünenberger, S. J.; Marrink, S. J. Martini extension to carbohydrates. *J. Chem. Theory Comput.* **2009**, *5*, 3195−3210.
(16) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. Multiscale modeling of emergent materials: biological and soft matter. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869−1892.
(17) Savelyev, A.; Papoian, G. A. Chemically accurate coarse graining of double-stranded DNA. *Proc. Natl. Acad. Sci. U. S. A* **2010**, *107*, 20340−20345.
(18) Bereau, T.; Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **2009**, *130*, 235106.
(19) Tschöp, W.; Kremer, K.; Han, O.; Batoulis, J.; Bürger, T. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polym.* **1998**, *46*, 61−74.
(20) Faller, R. Systematic Coarse Graining of Polymers and Biomolecules. In *Multiscale Modelling Methods for Applications in Materials Science*; Kondov, I., Sutmann, G., Eds.; Julich, Germany, September 16−20, **2013**; pp 135−150.
(21) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 090901.
(22) Ingólfsson, H. I.; López, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *WIREs Comput. Mol. Sci.* **2013**, *4*, 225−248.
(23) Brini, E.; Algaer, E. A.; Ganguly, P.; Rodriguez-Ropero, F.; van der Vegt, N. F. A. Systematic coarse-graining methods for soft matter simulations—a review. *Soft Matter* **2013**, *9*, 2108−2119.

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(24) Dans, P. D.; Zeida, A.; Machado, M. R.; Pantano, S. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.* **2010**, *6*, 1711−1725.
(25) Zeida, A.; Machado, M. R.; Dans, P. D.; Pantano, S. Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2012**, *86*, 021903.
(26) Darré, L.; Machado, M. R.; Dans, P. D.; Herrera, F. E.; Pantano, S. Another coarse grain model for aqueous solvation: WAT FOUR? *J. Chem. Theory Comput.* **2010**, *6*, 3793−3807.
(27) Darre, L.; Tek, A.; Baaden, M.; Pantano, S. Mixing atomistic and coarse grain solvation models for MD simulations: Let WT4 handle the bulk. *J. Chem. Theory Comput.* **2012**, *8*, 3880−3894.
(28) Darre, L.; Machado, M. R.; Pantano, S. Coarse-grained models of water. *WIREs Comput. Mol. Sci.* **2012**, *2*, 921−930.
(29) Gonzalez, H. C.; Darre, L.; Pantano, S. Transferable mixing of atomistic and coarse-grained water models. *J. Phys. Chem. B* **2013**, *117*, 14438−14448.
(30) Machado, M. R.; Dans, P. D.; Pantano, S. A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys. Chem. Chem. Phys.* **2011**, *13*, 18134−18144.
(31) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712−725.
(32) Dans, P. D.; Darre, L.; Machado, M. R.; Zeida, A.; Brandner, A. F.; Pantano, S. Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level. In *Advances In Bioinformatics and Computational Biology*; Setubal, J. C., Almeida, N. F., Eds.; Springer International Publishing: Gewerbestrasse, Switzerland, 2013; pp 71−81.
(33) Sterpone, F.; Melchionna, S.; Tuffery, P.; Pasquali, S.; Mousseau, N.; Cragnolini, T.; Chebaro, Y.; St-Pierre, J. F.; Kalimeri, M.; Barducci, A.; Laurin, Y.; Tek, A.; Baaden, M.; Nguyen, P. H.; Derreumaux, P. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chem. Soc. Rev.* **2014**, *43*, 4871−4893.
(34) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theor. Comp.* **2008**, *4*, 819−834.
(35) Maritan, A.; Micheletti, C.; Trovato, A.; Banavar, J. R. Optimal shapes of compact strings. *Nature* **2000**, *406*, 287−290.
(36) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A* **2001**, *98*, 10037−10041.
(37) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson−Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.
(38) Darden, T. A.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(39) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(40) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(41) Simmerman, H. K.; Jones, L. R. Phospholamban: Protein structure, mechanism of action, and role in cardiac function. *Physiol Rev.* **1998**, *78*, 921−947.
(42) Mortishire-Smith, R. J.; Pitzenberger, S. M.; Burke, C. J.; Middaugh, C. R.; Garsky, V. M.; Johnson, R. G. Solution structure of the cytoplasmic domain of phopholamban: Phosphorylation leads to a local perturbation in secondary structure. *Biochemistry* **1995**, *34*, 7603−7613.
(43) Mascioni, A.; Karim, C.; Zamoon, J.; Thomas, D. D.; Veglia, G. Solid-state NMR and rigid body molecular dynamics to determine domain orientations of monomeric phospholamban. *J. Am. Chem. Soc.* **2002**, *124*, 9392−9393.
(44) Pantano, S.; Carafoli, E. The role of phosphorylation on the structure and dynamics of phospholamban: a model from molecular simulations. *Proteins* **2007**, *66*, 930−940.
(45) Haenni, D.; Zosel, F.; Reymond, L.; Nettels, D.; Schuler, B. Intramolecular distances and dynamics from the combined photon statistics of single-molecule FRET and photoinduced electron transfer. *J. Phys. Chem. B* **2013**, *117*, 13015−13028.
(46) Glenner, G. G.; Keiser, H. R.; Bladen, H. A.; Cuatrecasas, P.; Eanes, E. D.; Ram, J. S.; Kanfer, J. N.; DeLellis, R. A. Amyloid. VI. A comparison of two morphologic components of human amyloid deposits. *J. Histochem. Cytochem.* **1968**, *16*, 633−644.
(47) Geddes, A. J.; Parker, K. D.; Atkins, E. D.; Beighton, E. "Cross-beta" conformation in proteins. *J. Mol. Biol.* **1968**, *32*, 343−358.
(48) Astbury, W. T.; Dickinson, S.; Bailey, K. The X-ray interpretation of denaturation and the structure of the seed globulins. *Biochem. J.* **1935**, *29*, 2351−2360.
(49) Makin, O. S.; Atkins, E.; Sikorski, P.; Johansson, J.; Serpell, L. C. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 315−320.
(50) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364−1368.
(51) Viana, I. F. T.; Dhalia, R.; Marques; Marques, E. T. A.; Lins, R. D. Influence of Scaffold Stability and Electrostatics on Top7-Based Engineered Helical HIV-1 Epitopes. In *Advances In Bioinformatics and Computational Biology*; Setubal, J. C., Almeida, N. F., Eds.; Springer International Publishing: Gewerbestrasse, Switzerland, 2013; pp 94−103.
(52) Berman, H. M.; Ten Eyck, L. F.; Goodsell, D. S.; Haste, N. M.; Kornev, A.; Taylor, S. S. The cAMP binding domain: An ancient signaling module. *Proc. Natl. Acad. Sci. U. S. A* **2005**, *102*, 45−50.
(53) Akimoto, M.; Zhang, Z.; Boulton, S.; Selvaratnam, R.; VanSchouwen, B.; Gloyd, M.; Accili, E. A.; Lange, O. F.; Melacini, G. A mechanism for the auto-inhibition of hyperpolarization-activated cyclic nucleotide-gated (HCN) channel opening and its relief by cAMP. *J. Biol. Chem.* **2014**, *289*, 22205−22220.
(54) Canaves, J. M.; Taylor, S. S. Classification and phylogenetic analysis of the cAMP-dependent protein kinase regulatory subunit family. *J. Mol. Evol.* **2002**, *54*, 17−29.
(55) Berrera, M.; Pantano, S.; Carloni, P. Catabolite activator protein in aqueous solution: A molecular simulation study. *J. Phys. Chem. B* **2007**, *111*, 1496−1501.
(56) Berrera, M.; Pantano, S.; Carloni, P. cAMP Modulation of the cytoplasmic domain in the HCN2 channel investigated by molecular simulations. *Biophys. J.* **2006**, *90*, 3428−3433.
(57) Thiru, A.; Nietlispach, D.; Mott, H. R.; Okuwaki, M.; Lyon, D.; Nielsen, P. R.; Hirshberg, M.; Verreault, A.; Murzina, N. V.; Laue, E. D. Structural basis of HP1/PXVXL motif peptide interactions and HP1 localisation to heterochromatin. *EMBO J.* **2004**, *23*, 489−499.
(58) Aasland, R.; Stewart, A. F. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Res.* **1995**, *23*, 3168−3173.
(59) Mendez, D. L.; Mandt, R. E.; Elgin, S. C. Heterochromatin protein 1a (HP1a) partner specificity is determined by critical amino acids in the chromo shadow domain and C-terminal extension. *J. Biol. Chem.* **2013**, *288*, 22315−22323.
(60) Jahn, R.; Scheller, R. H. SNAREs—Engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 631−643.
(61) Sudhof, T. C.; Rothman, J. E. Membrane fusion: Grappling with SNARE and SM proteins. *Science* **2009**, *323*, 474−477.
(62) Pantano, S.; Montecucco, C. The blockade of the neurotransmitter release apparatus by botulinum neurotoxins. *Cell. Mol. Life Sci.* **2014**, *71*, 793−811.
(63) Chen, X.; Tomchick, D. R.; Kovrigin, E.; Arac, D.; Machius, M.; Sudhof, T. C.; Rizo, J. Three-dimensional structure of the complexin/SNARE complex. *Neuron* **2002**, *33*, 397−409.
(64) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins* **2005**, *61*, 704−721.

Q

DOI: 10.1021/ct5007746
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX