Tesis de Doctorado en Ciencias Biológicas

Programa de Desarrollo de las Ciencias Básicas (PEDECIBA)
Área Biología
Sub área Biofísica

Universidad de la República

República Oriental del Uruguay | Julio 2012

MODELADO MOLECULAR DE PROCESOS RELACIONADOS A LA TRANSCRIPCIÓN DEL VIRUS VIH-1

Lic. Matías Machado (mmachado@pasteur.edu.uy)

Orientador: Dr. Sergio Pantano (spantano@pasteur.edu.uy)
Co-orientador: Dr. Pablo D. Dans (pdans@pasteur.edu.uy)

Grupo de Simulaciones Biomoleculares
Institut Pasteur de Montevideo
Mataojo 2020, CP 11400
Montevideo, Uruguay

Proyecto financiado por la Agencia Nacional de Investigación e Innovación (ANII) y La Comisión Sectorial de Investigación Científica (CSIC) – UdelaR.

RESUMEN

El Virus de la Inmunodeficiencia Humana (VIH) es un agente patógeno de gran impacto en la población humana. Este virus afecta células del sistema inmune integrándose de forma persistente en su genoma. Durante este estadío, llamado provirus, el VIH es capaz de usar la maquinaria celular para favorecer su replicación. Debido a que algunos provirus de la población infectiva pueden permanecer latentes, sin transcribir su genoma por un gran período de tiempo, las terapias existentes solo logran mitigar los efectos de la enfermedad pero no erradicar la infección. Comprender mejor los mecanismos implicados en la transcripción y represión del genoma viral es fundamental para el desarrollo de nuevas terapias contra este agente.

En el presente trabajo se aplicó el estado del arte en simulaciones de dinámica molecular para explorar aspectos atomísticos relacionados a los fenómenos de transcripción y represión. Dado que las técnicas de simulación están limitadas a sistemas moleculares y tiempos de simulación relativamente pequeños, se dedicó un tiempo significativo del trabajo de tesis al desarrollo de modelos simplificados que permitieran acceder a escalas temporales y espaciales biológicamente relevantes. De esta forma se generaron las herramientas necesarias para simular el comportamiento de la región promotora del VIH (80 pares de bases) en presencia y ausencia de la proteína de unión al elemento de regulación TATA. De este estudio se destaca la capacidad del ADN como medio para transmitir información en forma mecánica a través de su estructura. Por otro lado, se estudió a la Proteína Heterocromática 1 por ser un actor central en el establecimiento y mantenimiento del estado represivo de la cromatina que conduce a la latencia del virus. En este caso se evidenciaron determinantes estructurales en la interacción isoforma específica con la histona 3. Por último, partiendo del conocimiento estructural y bioquímico de varias proteínas que participan en la regulación del virus, se generó un modelo estructural del provirus de VIH-1 en estado de latencia. Este modelo permitió replantear algunos esquemas obtenidos de biología molecular, permitiendo de esta forma tender un puente entre la visión atomística y macroscópica de los procesos.

De esta manera, por medio del modelado molecular se logró cubrir un amplio espectro de sistemas relacionados a la transcripción del VIH.

En este trabajo de tesis se desarrollan en mayor detalle resultados que se encuentran aun en preparación. Los mismos serán enviados para su publicación en revistas arbitradas internacionales. La mayor parte del trabajo realizado ha sido publicado en los artículos citados a continuación:

1. Dans PD, Zeida A, **Machado MR**, Pantano S. (**2010**) A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory. Comput.* 6:1711-1725.
2. Zeida A, **Machado MR**, Dans PD, Pantano S (**2012**) Breathing, bubbling and bending: DNA flexibility from multi microseconds simulations. *Phys Rev E. 86: 021903 .* DOI: 10.1103/PhysRevE.86.021903
3. Darré L, **Machado MR**, Dans PD, Herrera FE, Pantano S. (**2010**) Another coarse-grain model for aqueous solvation: WAT four? *J. Chem. Theory. Comput.* 6: 3793-3807.
4. Dans PD, Darré L, **Machado MR**, Zeida A, Pantano S. (**2011**) Coarse grain potential: a model for DNA in implicit and explicit solvent. *En "A course on biomolecular simulations", ed. Jordi Villà-Freixa, Huygens Editorial*, en prensa.
5. **Machado MR**, Dans PD, Pantano S. (**2011**) A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys. Chem. Chem. Phys.* 13: 18134-18144.
6. **Machado MR**, Dans PD, Pantano S. (**2010**) Isoform-specific determinants in the HP1 binding to histone 3: insights from molecular simulations. *Amino Acids.* 5:1571-1581.

AGRADECIMIENTOS

Agradezco a la Agencia Nacional de Investigación e Innovación (ANII) por haber financiado la beca de Maestría y a la Comisión Sectorial de Investigación Científica (CSIC) de la UdelaR por la financiación de la beca de finalización de postgrado que cubrió la realización del Doctorado. Al Institut Pasteur de Montevideo, lugar en donde realice la tesis y el cual me financió para desarrollar este trabajo de investigación.

También agradezco a Sergio Pantano por haberme abierto las puertas de su laboratorio y brindado su apoyo y guía, también por su constante preocupación para que esta etapa de formación sea un trabajo debidamente remunerado. A Pablo Dans por los invalorables aportes y consejos como co-tutor de la Tesis. A los compañeros del Grupo de Simulaciones Biomoleculares con los cuales compartí muy buenas e innumerables experiencias: Leonardo Darré, Fernando Herrera, Ari Zeida, Humberto Gonzalez, Astrid Brandner y Sebastián Ferreira. A los amigos del Institut Pasteur de Montevideo que me ayudaron a integrarme a la vida del instituto. Al resto de los buenos amigos tanto de Facultad de Ciencias como de otros ámbitos que han estado presentes de una forma u otra animándome a seguir adelante.

Por último agradezco a mi familia por todo el apoyo brindado estos años, sin lo cual seria imposible haber logrado todas las metas propuestas. En particular a mis padres Darío y Carmen, muchas gracias. Agradezco especialmente a Mari por acompañarme, aconsejarme y soportarme durante este largo camino.

ÍNDICE DE CONTENIDO

INTRODUCCIÓN

El presente trabajo esta organizado de la siguiente forma: primero se expone una introducción general al problema y las herramientas de trabajo; posteriormente se plantean los objetivos; los capítulos y secciones subsiguientes cuentan con su propia introducción, metodología detallada, resultados, discusión y conclusiones. Solo se describen en detalle las partes del estudio que no han sido publicadas, enfatizando y resaltando la pertinencia e importancia de los trabajos publicados para el desarrollo de la Tesis. A modo de cierre del trabajo, se culmina con una serie de consideraciones globales.

El problema biológico

La infección con el Virus de la Inmunodeficiencia Humana (VIH) es una de las enfermedades más difundidas a nivel mundial. Según informes de UNAIDS (*United Nations Joint Programme on HIV/AIDS, http://www.unaids.org/*) y WHO (*World Health Organization, http://www.who.int/*) del año 2010 este virus es el responsable de una de las pandemias más destructivas de la historia, causando más de 25 millones de muertes en todo el mundo. Su incidencia en número de personas infectadas asciende aproximadamente a 33 millones en todo el mundo.

El VIH es un lentivirus envuelto de la familia *Retroviridae,* posee un genoma ARN monocatenario con polaridad positiva, compuesto por aproximadamente 10 mil nucleótidos. Se conoce la existencia de dos especies de VIH, denominadas VIH-1 y VIH-2 [1]. De las dos especies, VIH-1 es la más virulenta y es causa de la mayor cantidad de infectados a nivel mundial [2]. El “ciclo de vida” del virus puede racionalizarse y dividirse en varias etapas [1]. La primera etapa consiste en la unión de la partícula viral a la membrana de la célula blanco, esta interacción esta mediada por proteínas de la envoltura del virus (gp120 y gp41 que componen el complejo transmembrana Env, [3]) y de la célula hospedadora (CD4, CCR5, CXCR4 y otros). A continuación se produce la fusión de las membranas y liberación al medio intracelular del material genético del virus (ARN) y proteínas virales. El genoma viral, que ingresa como una simple hebra de ARN es transcripto a ADN de doble cadena por medio de la transcriptasa reversa del virus. Esta etapa es esencial para su posterior integración al genoma de la célula blanco. La integrasa del virus colabora en

esta última etapa [4]. Una vez embebido en el genoma eucariota, el provirus se comporta como un gen más del hospedador. Para favorecer tanto la síntesis de proteínas virales como de ARN, el virus cuenta con proteínas como Tat que recluta factores de transcripción hacia los promotores virales [5]. Las partículas maduras emergen por brotación de la membrana celular. No todas las partículas virales que infectan al hospedero siguen esta secuencia de etapas. Bajo condiciones particulares algunos provirus dejan de transcribir su genoma entrando en un estado de latencia [6]. Esta represión esta mediada principalmente por cambios en el estado de compactación de la cromatina en respuesta a señales epigenéticas [7,8]. Al no expresar proteínas, los provirus en estado latente logran escapar del sistema inmune y las terapias actuales, favoreciendo la persistencia de la infección en el hospedador [9]. En resumen, el estadío provirus del VIH-1 juega un rol muy importante durante la infección virus. En el se dan dos procesos de gran interés: transcripción y represión del genoma viral. Contribuir a entender los mecanismos asociados a ambos procesos es esencial para desarrollar nuevas terapias contra la enfermedad.

Los mecanismos de transcripción están íntimamente relacionados con diversos elementos de regulación presentes en la región promotora de los genes y con los factores de transcripción que estos unen [10]. Como ocurre en cualquier gen la transcripción del genoma viral debe ser un proceso regulado y sincronizado. Para ello colaboran varias proteínas actuando en *cis* y *trans* [8,11]. Una vía para modular la actividad de las proteínas es la unión de factores reguladores en regiones espacialmente distantes a los sitios efectores de las mismas pero que repercuten en su función. A este efecto se lo denomina alosterísmo [12] e involucra cambios en la conformación y/o dinámica de la proteína, los cuales son transmitidos a través de su estructura [13]. Actualmente se sabe que la fibra de ADN también es capaz de mediar efectos alostéricos como consecuencia de la unión de moléculas a su estructura [14]. En este contexto, es interesante preguntarse si existe una optimización en la composición de bases de las secuencias virales no solo para unir factores de transcripción, sino también para trasmitir señales entre elementos de regulación mediante efectos alostéricos y así contribuir a la coordinación de todo el proceso. La existencia de tal efecto podría ayudar a comprender la

relevancia de algunos polimorfismos observados en secuencias virales [15]. Para responder a esta pregunta se decidió centrar el estudio en la región próxima al inicio de la transcripción del VIH-1 (~80 pares de bases), donde se une la proteína TBP. Esta proteína genera una distorsión muy importante en el ADN [16], por lo que es un candidato ideal para estudiar las perturbaciones que genera sobre el entorno nucleotídico. Para abordar la complejidad y tamaño de los sistemas (más de 10^4 átomos) a estudiar se generaron modelos simplificados que permiten muestrear ventanas de tiempo biológicamente relevantes en estos procesos (microsegundos).

Por otra parte, los mecanismos de represión permiten que algunos provirus puedan permanecer transcripcionalmente inactivos generando reservorios celulares inaccesibles a los fármacos actualmente disponibles. Por lo tanto, la única cura definitiva a la enfermedad debe involucrar la erradicación de tales reservorios [9]. Para ello es central estudiar la estructura de la cromatina y las proteínas que junto a señales epigenéticas modifican su organización [7,17]. Entre las proteínas involucradas, se observó que la Proteína Heterocromática 1 (del inglés *heterochromatin protein 1*, HP1) juega un papel central en el VIH al favorecer el estado compacto de la cromatina y con esto contribuir a reprimir el genoma viral [18,19]. Trabajos recientes señalan a HP1 como un posible blanco molecular en nuevas terapias contra el VIH-1 al mediar tanto la interacción con la histona 3 (H3), componente de los nucleosomas que forman la cromatina, como el reclutamiento de metiltransferasas que modifican los nucleosomas con señales epigenéticas de silenciamiento [6]. Curiosamente se observó que diferentes isoformas de HP1 pueden actuar en estados de latencia o activación del genoma proviral, donde el cambio de isoforma esta mediado por modificaciones epigenéticas en H3. De aquí surge la interrogante sobre los factores estructurales que pueden determinar las distintas interacciones entre H3 e isoformas de HP1. Sin embargo, la HP1 es solo una pequeña parte del complejo rompecabezas de los posibles mecanismos responsables del estado de latencia del genoma proviral. Otras proteínas y factores de transcripción celulares contribuyen a mantener la represión viral [20], de ellos se conocen algunos fragmentos de su estructura pero nada sobre su orientación espacial y organización mesoscópica en el conjunto del sistema. Por lo tanto, para ampliar

nuestro conocimiento sobre la organización macromolecular espacial se integraron los datos disponibles en la bibliografía y las estructuras experimentales depositadas en el PDB en un modelo estructural del provirus de VIH-1 en estado de latencia.

Técnicas de modelado molecular

En el trabajo de tesis se emplearon principalmente técnicas de modelado y simulación molecular. Estas herramientas validadas con experimentos de biología molecular ofrecen una alternativa adecuada para obtener información estructural sobre estos sistemas. A continuación se describen brevemente los fundamentos teóricos detrás de las técnicas empleadas. Por mayores detalles en sobre estas herramientas se recomienda recurrir a bibliografía especializada [21].

Dinámica molecular

La Dinámica Molecular (DM) es una técnica de simulación en la cual se resuelven las ecuaciones del movimiento de Newton para describir el comportamiento temporal de un sistema de partículas. Típicamente las partículas corresponden a átomos.

La segunda ley de Newton se puede escribir como:

$$\vec{F}_i = m_i \left(\frac{d^2 \vec{r}_i}{dt^2} \right) \quad (1)$$

Conociendo las fuerza $\vec{F}_i$ que actúan sobre un átomo *i* de masa m_i es posible determinar las nuevas posiciones $\vec{r}_i$ de la partícula para un cierto paso de tiempo *dt*. En métodos de DM la Ecuación 1 se resuelve usando algoritmos de integración numérica. Un ejemplo de integrador es el algoritmo de Verlet [22], el cual calcula las posiciones y velocidades $\vec{v}_i$ para un tiempo $(t + \Delta t)$ como:

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \Delta t \, \vec{v}_i(t) + \frac{\Delta t^2}{2\mathrm{m}} \vec{F}_i(t) \quad (2)$$

$$\vec{v}_i(t + \Delta t) = \vec{v}_i(t) + \frac{\Delta t}{2\mathrm{m}} [\vec{F}_i(t) + \vec{F}_i(t + \Delta t)] \quad (3)$$

Para resolver las Ecuaciones 2 y 3 es necesario contar a un tiempo *t* con la posición $\vec{r}_i(t)$, la velocidad $\vec{v}_i(t)$, la fuerza $\vec{F}_i(t)$ y definir el paso de tiempo Δt para la integración. Las velocidades iniciales (*t* = 0) se obtienen a partir de una distribución de velocidades de Maxwell-Boltzmann. En las posteriores integraciones estas se obtendrán como resultado de la Ecuación 3. La fuerza sobre cada átomo del sistema se calcula a partir del cambio en la energía potencial respecto a las nuevas posiciones:

$$\vec{F}_i = -\left(\frac{dU}{d\vec{r}_i} \right) \tag{4}$$

Es posible expresar la energía potencial total *U* de un modo simple como una sumatoria de términos que representan distintas contribuciones. Una forma comúnmente empleada en programas de simulación molecular es:

$$U = \sum_{enlaces} k_b (r_{ij} - r_{eq})^2 + \sum_{ángulos} k_\theta (\theta_{ij} - \theta_{eq})^2 + \sum_{diedros} \frac{V_k}{2} (1 + \cos(n\omega - \gamma)) + \sum_{no-enlazante} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\} \tag{5}$$

Donde las sumatorias se aplican a todo par $i < j$ de átomos. El primer término modela el enlace covalente como un potencial armónico que depende de la distancia r_{ij}. Los parámetros que describen esta interacción son la constante de fuerza k_b y la posición de equilibrio r_{eq}. La contribución en los ángulos se calcula de forma análoga a los enlaces, empleando las constantes k_θ y θ_{eq}. Los diedros se representan usando funciones periódicas. La contribución a la energía potencial del ángulo torsional (ω) depende de la periodicidad (*n*), la fase (γ) y la amplitud de la barrera torsional (V_k). Por último las interacciones no enlazantes de corto y largo alcance están representadas por los potenciales de van der Waals (vdW) y Coulomb respectivamente. En éstos juega un rol muy importante el diámetro de vdW (σ), la profundidad del pozo de energía (ε) y la carga de cada átomo (q_i, q_j). La constante ε_0 es la permitividad en el vacío. Es importante notar que mientras los primeros tres términos computan interacciones entre vecinos cercanos, de 2 a 4 átomos conectados por enlaces, la componente no enalzante incluye idealmente al

resto de los átomos del sistema. Los parámetros de cada termino se definen para tipos de átomos considerando su naturaleza química y grupo funcional en que se encuentra. De este modo no solo se considera al átomo sino también al entorno en el que se encuentra. De esta forma es posible diferenciar, por ejemplo, un átomo de Carbono en un anillo aromático o en un grupo carboxilo al usar tipos de átomos diferentes para cada uno. Al conjunto de parámetro, tipos de átomos y componentes de la energía se le denomina campo de fuerza. Dentro de los campos de fuerza más populares para la simulación de sistemas biológicos se encuentran AMBERff [23], CHARMM [24], GROMOS [25] y OPLS [26].

El paso de tiempo empleado para integrar las ecuaciones de movimiento se elige de forma tal que permita conservar la energía del sistema y genere un correcto muestreo del espacio de fases impidiendo solapamiento entre las partículas. Para cumplir estas condiciones se emplean pasos de tiempo de $1x10^{-15}$ s (1 femtosegundo). Esto permite muestrear correctamente vibraciones tan rápidas como las presentes en enlaces de átomos pesados (N, C, O, etc.) con hidrógeno. Por lo tanto simular sistemas moleculares por tiempos de 10^{-9} s a 10^{-6} s (nanosegundos a microsegundos) requiere de efectuar 10^6 a 10^9 pasos de simulación (integración de las ecuaciones del movimiento), evaluando las fuerzas cada vez. Esto demanda un gran tiempo de cálculo, en particular para sistemas biológicos grandes [27,28].

Si bien los elementos mencionados constituyen el corazón de toda DM, existen otros detalles prácticos importantes a mencionar. Para realizar una simulación dentro de un ensemble estadístico adecuado es necesario cumplir con ciertas condiciones. Las condiciones más comunes se comentan a continuación. El número de partículas permanece constante durante toda la simulación. Es necesario acoplar el sistema a un termostato para mantener la temperatura constante. Existen distintos métodos que permiten controlar esta propiedad (termostatos de Nosé-Hoover [29,30], Langevin, etc.). La forma más simple implica escalar las velocidades, un ejemplo de como hacerlo es el termostato de Berendsen [31], donde las velocidades de cada partícula se escalan por el factor λ luego de n_{TC} pasos de simulación:

$$\lambda = \left[1 + \frac{n_{TC} \Delta t}{\tau_T} \left(\frac{T_0}{T\left(t - \frac{1}{2}\Delta t\right)} - 1 \right) \right]^{\frac{1}{2}} \quad (6)$$

Este factor depende de la temperatura observada (T) y la de referencia (T_0), así como del factor τ_T cuyo valor es cercano a la constante térmica de acoplamiento temporal. Otras propiedades importantes a controlar son la presión y el volumen. Ambas propiedades se regulan mediante el tamaño de la celda a simular, es decir el espacio total en el que se define el sistema. Los barostatos más usados son Berendsen [31] y Parrinello-Rahman [32,33]. Con estas herramientas es posible simular el sistema en condiciones de número de partículas, temperatura y presión constantes (NTP, ensemble isotérmico-isobárico) o volumen constante en lugar de presión (NTV, ensemble canónico). Lo cual permite calcular propiedades termodinámicas a partir del promedio temporal de configuraciones del sistema usando los conceptos de la termodinámica estadística.

En los sistemas biológicos el solvente cumple un rol muy importante tanto para mediar interacciones como en el plegamiento y estabilización de macromoléculas. Por lo tanto es imprescindible considerar su contribución, lo que se puede hacer tanto explícita como implícitamente. La descripción explícita del solvente implica tener dentro del sistema de estudio las moléculas de agua e iones. Se han desarrollado varios modelos para representar las moléculas de agua. Los modelos más simples y ampliamente utilizados son TIP3P [34] y SPC [35]. En ambos casos, cada molécula de agua es representada por un oxígeno unido de forma rígida a dos átomos de hidrógeno. Cada átomo posee una carga parcial que contribuye a la energía potencial de Coulomb, pero sólo el oxígeno cuenta con parámetros que contribuyen al potencial de vdW.

Por otro lado, los modelos de solvente implícito proporcionan una forma alternativa y eficiente de representar los efectos electrostáticos de las moléculas del solvente, y a su vez permiten ahorrar buena parte de los cálculos necesarios para describir de forma exacta (explícita) la solución acuosa circundante. Estos métodos se basan en capturar la influencia de las moléculas

de solvente sobre el soluto por medio de la energía libre de solvatación (ΔG_{solv}). Esto implica calcular el trabajo reversible de transferir el soluto desde el vacío al medio acuoso manteniendo fija su configuración. Dado que la dinámica del solvente es típicamente mucho más rápida que la difusión conformacional del soluto, el entorno acuoso puede ser descripto como un medio continuo. Si bien esta es solo una posibilidad, es una aproximación muy utilizada [36]. De los varios modelos de continuo desarrollados [37] uno de los más utilizados es el llamado modelo de Born Generalizado (GB) [38-40]. En este esquema la energía libre de solvatación se descompone en contribuciones no polares y electrostáticas:

$$\Delta G_{solv} = \Delta G_{np} + \Delta G_{elect} \tag{7}$$

La componente no polar (ΔG_{np}) se estima mediante la superficie accesible al solvente y una constante de proporcionalidad. Mientras que el término electrostático (ΔG_{elect}) se aproxima como:

$$\Delta G_{elect} = \left(1 - \frac{1}{\epsilon}\right) \sum_{ij} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j \exp\left(\frac{-r_{ij}^2}{4 b_i b_j}\right)}} \tag{8}$$

En esta expresión, los elementos más importantes son la distancia r_{ij} entre cada par de átomos del soluto, las cargas (q_i, q_j) de los mismos, y los radios de efectivos de Born (b_i, b_j). Estos últimos representan la distancia entre un átomo particular y el entorno esférico efectivo del dieléctrico. Mientras más exacto sea el radio de Born mejor será la descripción de la componente electrostática.

Modelos simplificados

Dos factores limitan computacionalmente la escala temporal de los fenómenos a observar con DM. Esto son el tamaño del sistema, medido en número de partículas (N), y el paso de integración (dt) de las ecuaciones de movimiento. Como las fuerzas se evalúan mediante un potencial de pares y en particular la interacción electrostática es de largo alcance incluyendo todos los átomos del sistema, el orden de complejidad del algoritmo que computa todas las interacciones en un dt es O(N^2). El número total de iteraciones en una simulación será por tanto $N^{2*}(t/dt)$, donde el cociente entre el tiempo total a

simular (*t*) y d*t* es la cantidad total de pasos de DM. Para acelerar el cálculo *N* debe ser pequeño y/o d*t* grande. Haciendo uso de algunos algoritmos es posible reducir el costo computacional en la evaluación de la fuerza a O(*N*logN*) [41]. Sin embargo, como se vio en la sección anterior el d*t* sigue siendo restrictivo. Esto hace que sistemas biológicos con $N \sim 10^6$ no puedan ser simulados por escalas de tiempo mayores a decenas de nanosegundos (10^{-9}s) con el poder de cálculo disponible en la actualidad [28]. Esta característica restringe el estudio de fenómenos biológicamente relevantes, como ser cambios conformacionales, que ocurren en escalas de tiempo mayores a los nanosegundos.

Una posible vía para ampliar la escala temporal de la DM es usar modelos simplificados. El concepto subyacente a esta estrategia implica desarrollar modelos de menor complejidad pero que representan con suficiente aproximación al sistema original en alguna propiedad de interés. No existe una forma única de construir modelos simplificados, es por ello que la literatura es muy vasta en este tema (por una completa revisión leer [42]). Existen ejemplos para todos lo tipos de biomoléculas: proteínas, ácidos nucleicos, lípidos. Sin embargo, al perder grados de libertad en el sistema siempre se debe elegir que característica reproducir mejor, puesto que será imposible tener la precisión deseada simultáneamente en todas las propiedades. Por lo tanto, rara vez un modelo simplificado es aplicable de forma general al estudio de cualquier sistema. En este punto es válido mencionar el campo de fuerza MARTINI [43,44], el cual posee parámetros para varias biomoléculas exceptuando los ácidos nucleicos.

Una manera de simplificar la topología de una molécula o residuo es agrupar con cierto sentido físico-químico átomos en centroides efectivos de interacción. Por ejemplo, el modelo de MARTINI simplifica el grupo benceno por tres centroides, cada uno de los cuales representa aprox. 2 átomos de Carbono y 2 de Hidrógeno. Estos puntos de interacción (centroides) se ubican en el centro de masa de dos grupos CH de la molécula original. La masa total del benceno se distribuye de forma equivalente entre cada centroide. En este modelo simplificado del benceno los centroides están enlazados entre si, lo que da un aspecto triangular a una topología originalmente hexagonal. Este ejemplo

sencillo ilustra como se ve alterada la conectividad y rugosidad de la superficie molecular. Para reproducir correctamente las propiedades del benceno los parámetros de interacción deben cambiar respecto a su contraparte atómica original. Por ejemplo el radio de van der Waals debe aumentar para dar cuenta del mayor volumen excluido de cada centroide, la carga se debe ajustar, etc. Un dato relevante es que este modelo simplificado reduce 75% el tamaño del sistema. Como los centroides poseen más masa y los enlaces vibran con menor frecuencia, es posible aumentar el paso de integración a 50fs. De este modo tanto *N* como d*t* cambian de forma favorable, y la velocidad de cálculo con el modelo simplificado aumenta varios ordenes de magnitud respecto de la misma simulación en la versión con detalle atómico.

Durante el desarrollo de esta tesis se contribuyo en el diseño de un modelo simplificado de ADN (ver Artículos 1 al 4 sección Publicaciones [45-48] para mayores detalles).

OBJETIVOS GENERALES

Dilucidar factores moleculares relacionados a los mecanismos de transcripción y represión del VIH-1. Además de su relación con la infección por VIH-1, se espera contribuir al conocimiento básico en temas relacionados con la regulación de la transcripción en regiones promotoras. El trabajo también pretende contribuir a la creación de nuevas herramientas que permitan el abordaje de sistemas cada vez más grandes y complejos en el área del modelado molecular.

Objetivos específicos

Mecanismos moleculares implicados en la transcripción del VIH-1

1. *Desarrollo de modelos simplificados para ácidos nucleicos.*
2. *Estudio de la región promotora del VIH-1 y su interacción con TBP.*

Mecanismos asociados a la represión y estado de latencia del VIH-1

1. *Caracterización de determinantes estructurales en la interacción isoforma específica de HP1 con la histona 3.*
2. *Ensamblado molecular del provirus de VIH-1 en estado de latencia.*

CAPÍTULO 1: Mecanismos moleculares implicados en la transcripción del VIH-1

La estructura del genoma de VIH-1 se organiza en tres grandes regiones: dos regiones reguladoras iguales en secuencia de aprox. 600 nucleótidos cada una, las que se ubican respectivamente en los extremos 5' y 3' del genoma. A estas regiones se las denomina LTR (del inglés: *Long Terminal Repeat*) y son la llave para activar o reprimir la expresión viral. Por otro lado, existe una región central que contiene los genes gag, pol, tat y env entre otros, los cuales codifican para proteínas virales [49]. Al igual que ocurre en otros retrovirus la región 5'LTR controla la expresión del genoma viral. La estructura fina del 5'LTR muestra diversas cajas de regulación en las que se presentan elementos TATA, NF-κb, SP1 entre otros (Figura 1.0.1) [50]. Estos elementos están involucrados en la unión a factores de transcripción que modifican la expresión del genoma viral.

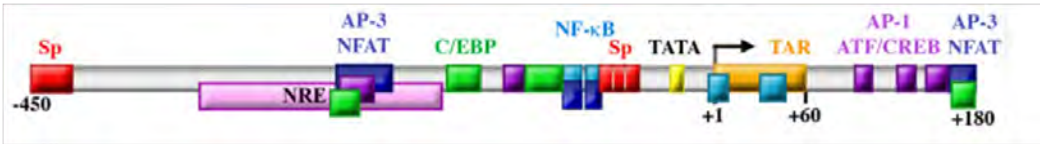

FIGURA 1.0.1 | Distintos elementos de regulación presentes en la región 5'LTR del genoma de VIH-1. Figura adaptada de [11].

Durante el estadío provirus, el genoma de VIH-1 se comporta como un gen más de la célula hospedadora. Así como ocurre en cualquier gen, su expresión estará mediada por un diálogo entre distintos elementos reguladores muchas veces distantes en secuencia [10]. En este diálogo, la unión de proteínas a secuencias reguladoras y la posterior interacción proteína-proteína es una muy importante vía para la transmisión de señales que desencadenan el proceso [51,52]. Sin embargo no está claro si el ADN de la región promotora solo participa como una plataforma de unión a proteínas o también podría canalizar señales mecánicas o alostéricas a través de su estructura [14,53]. Lejos de ser rígido, el ADN posee gran dinámica y flexibilidad, mostrando cambios de curvatura y torsión, apertura y cierre de surcos mayor y menor [54], así como eventos de desapareamiento de bases [55]. Por lo tanto, es posible pensar que cambios de estos patrones estructurales puedan mediar señales entre elementos de regulación no inmediatamente adyacentes. Entender los efectos

ejercidos por las interacciones ADN-proteína sobre la dinámica y flexibilidad intrínseca del ADN es clave para comprender mejor el funcionamiento de las secuencias que controlan la transcripción del VIH. Los resultados generados también tienen validez en un contexto más general de la transcripción génica e implicaciones en el diseño de promotores.

Las principales preguntas que busca responder este capítulo son: ¿Cuál es y cómo se afecta la dinámica y flexibilidad del ADN en la región promotora del virus VIH-1 al actuar en *cis* factores de transcripción? ¿Puede el ADN actuar como una vía molecular de comunicación entre distintos elementos de regulación o es meramente una plataforma pasiva para la unión a proteínas?

Usar herramientas computacionales para dar respuesta a estas preguntas ofrece una gran ventaja en relación al costo experimental y nivel de detalle molecular de los fenómenos a estudiar. Abarcar todo el amplio espectro de factores de transcripción excede el trabajo de la presente tesis doctoral, por lo que se acotó el estudio a uno solo. Se eligió a la proteína de unión al elemento TATA (TBP) como modelo, ya que representa uno de los primeros eventos de enlace proteína-ADN en la cadena de interacciones que dan lugar a la transcripción mediada por la RNA polimerasa II [56,57]. La TBP genera cambios conformacionales dramáticos sobre el ADN que implican la formación de quiebres en de la región de unión así como el pasaje a la forma A [58], por lo tanto se espera que estos efectos impacten en la dinámica de regiones próximas del ADN. Sumado a esto, la existencia de información estructural experimental para complejos ADN-TBP [59-61] posibilita su uso en el modelado de regiones promotoras. Como las propiedades a evaluar están asociadas al comportamiento temporal del ADN, se emplearon técnicas de dinámica molecular. En este marco, la gran limitante es el tamaño de los sistemas a simular, lo que dificulta el acceso a escalas temporales relevantes para los fenómenos a observar. Por lo tanto, lo primero que se hizo fue desarrollar modelos simplificados que luego fueron utilizados para simular los sistemas de interés biológicos.

Las siguientes secciones detallan el desarrollo de los modelos simplificados y su posterior aplicación al estudio de la región promotora del VIH-1 en ausencia o presencia de TBP.

Desarrollo de modelos simplificados para ácidos nucleicos

Se han propuestos varios modelos simplificados para estudiar la dinámica de los ácidos nucleicos. Sus resoluciones varían desde un centroide para representar cientos de pares de bases hasta aquellos que pueden distinguir entre residuos. Se limitaran los comentarios a estos últimos puesto que describir tanto torsiones o curvaturas locales como eventos de fusión de la doble hebra requiere considerar interacciones específicas entre pares de bases. Uno de los modelos más simples propuesto emplea un centroide por cada nucleótido [62]. Para mantener la conformación correcta del ADN es necesario definir una red de interacciones entre varios centroides cercanos. Modelos más detallados incluyen parte del esqueleto fosfato del ADN con lo cual pueden describir fácilmente su conformación. Un ejemplo es el modelo de Knotts *et al.* [63] que emplea 3 centroides por nucleótido. Un centroide representa el grupo fosfato, otro al azúcar y el último a la nucleobase. Usando parámetros específicos para cada tipo de base (adenina, guanina, tirosina o citosina) se logran describir comportamientos secuencia dependiente en la temperatura de fusión de la doble hebra. El modelo propuesto por Savin *et al.* agrega un poco más de detalle [64]. En él se utiliza un centroide para describir el grupo fosfato, dos para el azúcar y 3 para la nucleobase. De estos últimos solo 2 centrodes se emplean para describir el apareamiento Watson-Crick, mientras que el tercero se ubica en la posición del átomo C7 en pirimidinas o C8 en purinas. Si bien el modelo es capaz de describir propiedades térmicas del ADN, los propios autores no lo recomiendan para estudiar eventos que involucran desapareamiento de bases.

Una característica común a todos los modelos descriptos es omitir o restringir el cálculo explícito de la electrostática. Es importante recordar que interacciones del tipo ADN-proteína dependen del reconocimiento electrostático. Una mala descripción de esta contribución puede resultar en una incorrecta descripción de la física del problema. A su vez, para mitigar la falta de grados de libertad en los modelos, varios usan términos no estándar en la energía potencial. Esto limita el uso de estas aproximaciones a códigos desarrollados por los autores para simular su modelo en particular.

Durante este trabajo se desarrolló un modelo simplificado para ADN que incluye la electrostática explícita y que puede ser usado en programas estándar de DM. El modelo propuesto reduce la complejidad de los nucleótidos de aprox. 30 átomos a 6 centroides, (ver Figura 1.1.1, Artículo 1 sección Publicaciones, [45]). Esta simplificación reduce un 80% el tamaño del sistema y permite incrementar un orden de magnitud el paso de integración de la simulación llegando a 20fs, con lo cual el cálculo es 2400 veces más rápido respecto a un sistema atomístico idéntico en composición. La topología y parámetros del modelo permiten preservar la identidad fisicoquímica en las interacciones más relevantes de la forma B del ADN: enlaces de hidrógeno y apilamiento entre bases (Figura 1.1.1). Con esto, el modelo logra reproducir no solo propiedades helicoidales de la doble hebra de ADN, sino también su temperatura de fusión en función de la longitud, secuencia y fuerza iónica del medio, así como la dinámica de apareamiento entre la bases que da lugar a la formación de "burbujas" (zonas de varios pares de base desapareados) en el ADN [45]. Estas propiedades tienen un rol importante en elementos de regulación de secuencias promotoras (ejemplo caja TATA), como se detalla en la próxima sección.

A mayor escala, el ADN se comporta como un polímero lineal capaz de curvarse y enrollarse en función del largo de su secuencia. En este sentido, el modelo de ADN simplificado puede reproducir la longitud de persistencia del polímero (50 pares de bases) así como la formación de quiebres espontáneos en la doble hebra (Artículo 2 sección Publicaciones, [47]). La capacidad de muestreo del modelo permitió comparar la flexibilidad intrínseca del ADN con datos experimentales. De forma notable, se observó que el ADN por si solo es capaz de alcanzar una distribución de curvaturas similar a la de complejos ADN-proteína depositados en la base de datos *Protein Data Bank* (Artículo 2 sección Publicaciones, [47]). Esto implica que las conformaciones de unión pueden ser estados accesibles para el ADN previo a la interacción con factores de transcripción, favoreciendo la formación del complejo final.

Por último, dado que el modelo incluye explícitamente la electrostática a largo alcance otros efectos más finos como ser la hidratación y la interacción

con iones también se reproducen correctamente (ver Artículo 3 sección Publicaciones, [46]).

De esta forma se demostró que el modelo de ADN simplificado es capaz de reproducir una serie de propiedades locales y globales de la biomolécula. Una revisión detallada del alcance y limitantes del modelo puede obtenerse del Artículo 4 sección Publicaciones, [48].

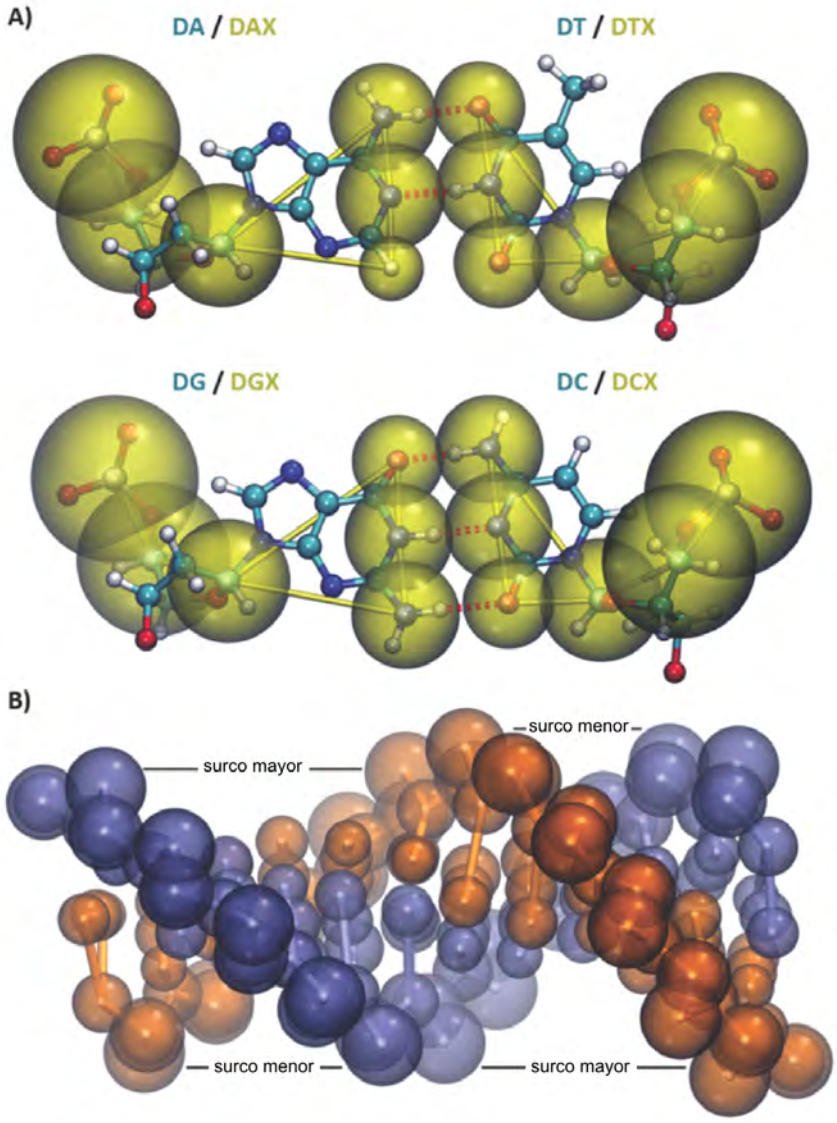

FIGURA 1.1.1 | Modelo simplificado de ADN. A) Correspondencia entre pares de base atomísticos y simplificados. Los seis centroides empleados para describir cada base en el modelo simplificado se representan con esferas de van der Waals amarillas. Como indica la figura los centroides ocupan las mismas posiciones de átomos particulares en las bases atomísticas. La conectividad entre los centroides se indica con varillas amarillas. El apareamiento de bases esta mediado por interacciones no-enlazantes. La posibilidad de formar enlaces de hidrógeno de Watson-Crick se muestra en punteado. B) Doble hebra de ADN en conformación B descripto a nivel simplificado. La estructura se muestra usando la misma representación empleada en (A). Cada hebra de ADN se muestra con un color distinto. La superficie molecular descripta por el modelo simplificado se asemeja al modelo atomístico. Es claro apreciar en la topología los surcos mayores y menores del ADN.

Existen situaciones en las cuales es conveniente o necesario incluir el detalle atómico para capturar todo el efecto de cierto proceso. Es así que se han desarrollado estrategias denominadas multiescala, en las cuales, una región pequeña del sistema es modelada a un nivel fino y otra mayoritaria a un nivel grueso (simplificado) [65]. Esto permite lograr un compromiso conveniente entre costo computacional y nivel de detalle en la simulación por DM. Con el objetivo de estudiar interacciones ADN-proteína dentro del contexto de una región promotora se desarrolló un modelo multiescala para simular ADN (ver Figura 1.1.2 y Artículo 5 sección Publicaciones, [66]).

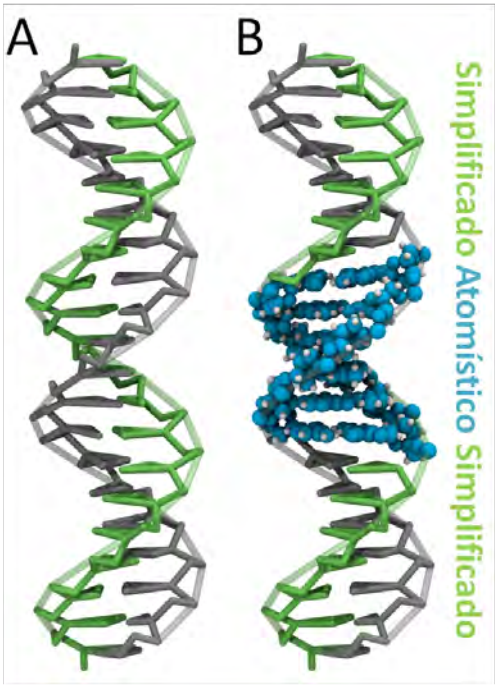

FIGURA 1.1.2 | Comparación entre el modelo simplificado y multiescala de un polímero de 20 pares de bases estudiado en el Artículo 5 sección Publicaciones, [66]. A) Modelo simplificado. Las estructura del modelo está representadas por varillas. Cada hebra de ADN se representa de un color distinto. Los tubos transparentes remarcan el esqueleto fosfato del ADN, pero no son parte de la topología del modelo. B) Sistema multiescala, en el cual la región central de seis pares de bases es descripta a nivel atomístico (esferas azules y blancas que corresponden a átomos pesados e hidrógeno). El contexto nucleotídico se representa a nivel simplificado. La leyenda sobre el margen derecho indica el límite de cada región.

Este esquema permite combinar el modelo simplificado de ADN, previamente introducido, con una descripción a nivel atomístico en alguna región del polímero. Las características más salientes del modelo son: simplicidad, compatibilidad y velocidad. Pocos parámetros son requeridos para combinar ambos niveles de descripción con escasa o nula perturbación en la frontera. Tanto los parámetros estructurales como las interacciones de larga

distancia se encuentran bien balanceadas. Según el análisis de componentes principales, el modelo no modifica el comportamiento dentro de cada región ni el de la macromolécula en su conjunto. El estudio realizado sobre una horquilla de ADN en el cual se logran reproducir resultados experimentales secuencia dependientes, muestra la flexibilidad y potencialidad de su aplicación (ver Artículo 5 sección Publicaciones, [66]). En este contexto, extender su uso a sistemas ADN-proteína no implica ningún esfuerzo adicional. La implementación del modelo es directa en paquetes estándar de simulación (ejemplo: AMBER [23], GROMACS [67]). Con esta metodología el costo computacional proviene mayormente de la región atomística.

El conjunto de herramientas desarrolladas no solo permite aplicar el estado del arte en técnicas de dinámica molecular para estudiar el comportamiento intrínseco de regiones promotoras, sino que posibilita explorar el efecto que tiene la unión de factores de regulación sobre estas secuencias.

Estudio de la región promotora del VIH-1 y su interacción con TBP

La posibilidad de que el ADN medie efectos alostéricos fue observada hace más de 40 años y hoy en día existe fuerte evidencia de su factibilidad [14,53]. Si embargo el peso que tiene este efecto en los mecanismos de transcripción es poco claro. En algunos potenciadores podría ser tan crucial como la interacción proteína-proteína [14]. Pese a esto todavía se conoce poco del detalle atómico que pueda dar origen a la transmisión de información a través de la estructura del ADN. En particular resta por entender más sobre los patrónes estructurales que se ven alterados durante la unión a proteínas y cuanto podrían extenderse en secuencia las perturbaciones.

Dentro de las propiedades estructurales más salientes del ADN se pueden mencionar las torsiones en el esqueleto fosfato, que dan lugar a enrollamientos o desenrollamientos de la doble hebra, apertura o cierre en los surcos y apareamiento de bases. Estos fenómenos son importantes en los mecanismos de lectura indirecta del ADN por parte de las proteínas [16], por lo que entender como se ve afectada su dinámica en distintos contextos puede dar más pistas sobre los factores que influyen en la regulación de la transcripción. Evaluar estas propiedades estructurales constituye un reto desde el punto de vista experimental, sin embargo es abordable con técnicas de dinámica molecular.

Como sistema de estudio se empleó la región promotora del VIH-1. Conocer los mecanismos implicados en la transcripción de este virus genera gran interés tanto a nivel básico como aplicado. Los resultados se contrastaron con simulaciones de un promotor denominado SP1 (*del inglés: Super Core Promoter 1*, [68]). El SCP1 es una secuencia ingenierizada a partir de fragmentos de otros promotores, con la cual se logra una alta tasa de transcripción *in vitro* e *in vivo*. Para evaluar el impacto que podrían tener pequeños cambios en secuencia se estudiaron cuatro mutaciones en SCP1 que reducen 80% su nivel de transcripción [69]. Estas mutaciones, representadas en el promotor llamado m1SCP1, no alteran la unión del ADN a factores de transcripción. Nuestro conocimiento previo de los sistemas SCP1 y m1SCP1 nos permite saber que el modelo simplificado de ADN, cuyo desarrollo se comenta en la sección anterior y en el Artículo 1 sección

Publicaciones [45], es suficientemente sensible tanto para identificar comportamientos secuencia dependiente en elementos de regulación (ejemplo TATA) como para discriminar efectos de mutaciones puntuales (Artículo 2 sección Publicaciones, [47]).

En este trabajo se pretende explorar el efecto que genera el enlace de una proteína al ADN sobre las propiedades dinámicas y estructurales del entorno próximo a la región de interacción. Para ello se empleó a la proteína de unión al elemento TATA (TBP) como paradigma de interacción ADN-proteína. Según se ha medido experimentalmente, TBP es capaz de inducir cambios notables de flexibilidad en la fibra de ADN [70]. También es muy bien conocida la estructura atómica de esta proteína, para la cual se han resuelto varios complejos con ADN, en alguno de los cuales participan otros factores de transcripción [56].

Metodología

Se estudiaron siete sistemas moleculares (Figura 1.2.1):

(i) Sistema S_{SCP1}: modelo simplificado del promotor SCP1, cuya secuencia en pares de base se extiende desde -36 a +45 [68]. Para emular el sistema de expresión experimental se empleó la secuencia del plásmido pUC119 en torno a los sitios de restricción *PstI* y XbaI. De este modo se emula la continuidad del promotor evitando efectos de borde. El polímero resultante contiene la siguiente secuencia: 5'GCATGCCTGCAGGTACT**TATATAA**GGGGGTGGGGGCGCGTTC GTCCTC(A)GTCGCGATCGAACACTCGAGCCGAGCAGACGTGCC TACGGACCGTCTAGAGGATCC3'. En negrita se señala la caja TATA, en sombreado la región de unión a TBP (pares de base a 7 Å de la proteína), en subrayado la secuencia de iniciadora (Inr) y entre paréntesis el sitio +1.

(ii) Sistema $S_{SCP1+TBP}$: La secuencia utilizada en este caso fue idéntica a la del sistema S_{SCP1} pero en lugar de una doble hebra de ADN en forma B canónica, se incluyo la distorsión estructural en la caja TATA. La conformación de los nucleótidos involucrados en el enlace a TBP fue forzada a mantenerse en una conformación equivalente a la observada en la estructura cristalográfica de complejo (código PDB

1C9B). De este modo se representó el efecto generado en el ADN por la unión a TBP.

(iii) Sistema S_{m1SCP1}: Ídem al sistema S_{SCP1} pero se incluyen las mutaciones propuestas por Alexandrov *et al.* [69]: T y C en posiciones -5 y -4 fueron mutadas por C y G respectivamente, A en posiciones +8 y +15 pasaron a ser G.

(iv) Sistema $S_{m1SCP1+TBP}$: Ídem al sistema S_{m1SCP1} pero incluye la distorsión estructural generada en el ADN por la unión a TBP, como se explicó para el sistema ii.

(v) Sistema S_{VIH}: Modelo simplificado de la región promotora del virus VIH-1 (código GenBank: K03455), secuencia comprendida entre las bases 391 y 467 (posiciones -65 y +13 relativo al sitio +1): 5'GCGGGACTGGGGAGTGGCGAGCCCTCAGATCCTG**CATATAA**G CAGCTGCTTTTTGCCTGTACTG(G)GTCTCTCTGGTT3'. En negrita se señala la caja TATA, en sombreado la región de unión a TBP y entre paréntesis el sitio +1.

(vi) Sistema $S_{VIH+TBP}$: Ídem al sistema S_{VIH} pero incluye la distorsión estructural generada en el ADN por la unión a TBP, como se explicó para los sistemas ii y iv.

(vii) Sistema $S^*_{VIH+TBP}$: Modelo multiescala del sistema $S_{VIH+TBP}$. En este sistema tanto la región de unión a TBP en el ADN como la proteína son representados a nivel atomístico, mientras el resto del sistema se simula a nivel simplificado.

En todos los casos el procedimiento de construcción de los modelos se inició a partir de coordenadas cartesianas de estructuras que contienen todos los átomos. Los promotores sin TBP se generaron en la forma B canónica de ADN mediante la utilidad NAB de AMBER11 [23]. Este procedimiento se aplico para construir todo fragmento de ADN desnudo, es decir carente de proteínas unidas. Para generar los sistemas con TBP se dividió la secuencia promotora en tres fragmentos. El primer fragmento comprendió la secuencia desde el extremo 5' del promotor hasta la región de unión a TBP, pero sin incluir a esta última. La región de unión a TBP se modeló a partir de la estructura con código

en el PDB 1C9B, la cual contiene la proteína TBP humana y el factor TFIIB ligados a una secuencia de ADN de 18 pares de bases. De esta estructura se eliminaron las proteínas y se mutaron los nucleótidos en el molde de ADN para que coincidieran con la región del promotor a simular.

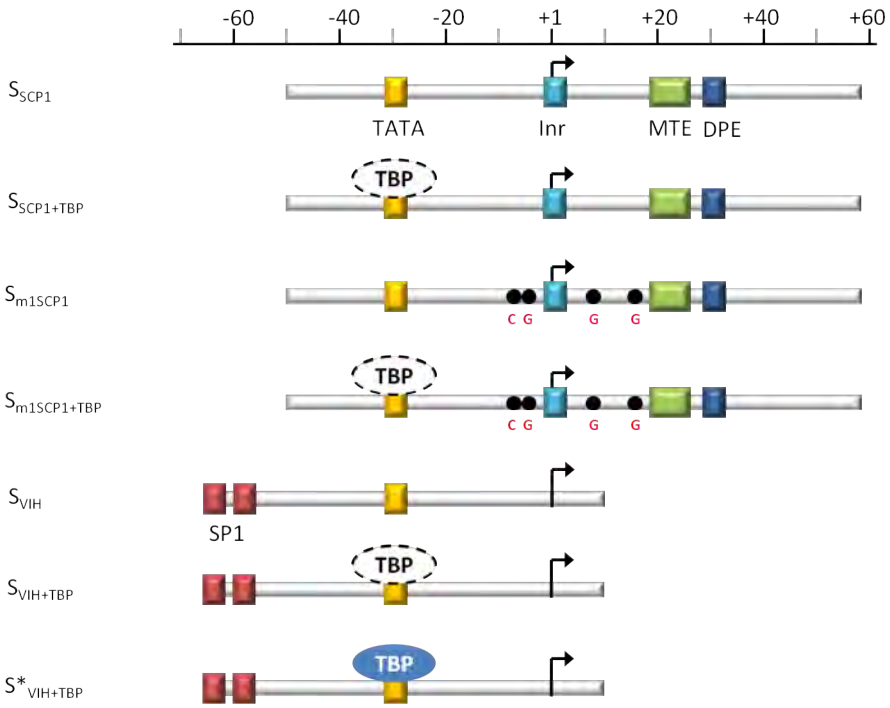

FIGURA 1.2.1 | Representación esquemática de los sistemas simulados. Las secuencias se numeran de forma relativa al sitio de iniciación de la transcripción (+1). Las cajas señalan elementos reguladores presentes en cada sistema como ser la secuencia TATA, la secuencia iniciadora (Inr), el motivo 10 (MTE), el elemento promotor corriente abajo (DPE) y la secuencia de unión a la proteína SP1. Los círculos negros en los sistemas S_{m1SCP1} y $S_{m1SCP1+TBP}$ indican mutaciones respecto al sistema S_{SCP1}. La representación punteada de TBP implica que su unión al ADN se modeló de forma implícita, mientras que la representación sólida en el sistema $S^*_{VIH+TBP}$ corresponde a la consideración explícita de la proteína.

Posteriormente se generó un último fragmento de ADN para completar la secuencia restante hacia el extremo 3' del promotor. Todos los fragmentos se diseñaron con cinco nucleótidos solapantes, esto permitió su ensamblado final mediante alineamiento estructural. Para mantener la continuidad en la cadena de ADN, se eliminaron los nucleótidos repetidos. La Figura 1.2.2 resume todo el procedimiento empleado en la construcción de los sistemas que presentan TBP.

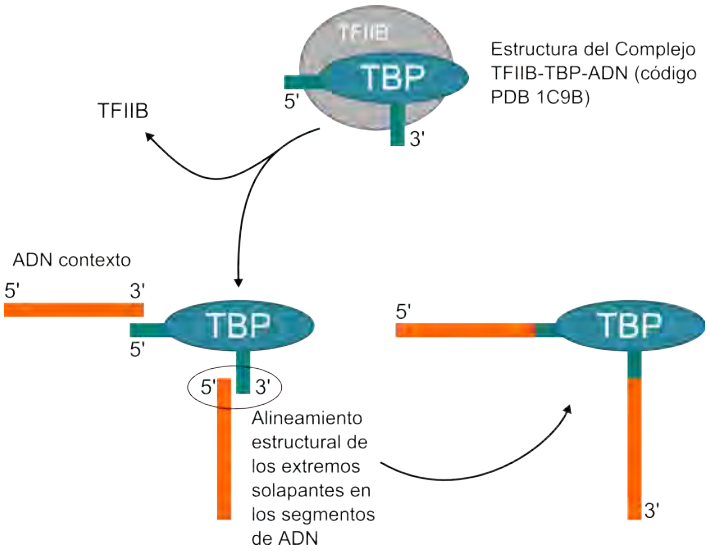

FIGURA 1.2.2 | Ensamblado de los sistemas promotores con TBP unido. Se muestra esquemáticamente el procedimiento por el cual se genera un modelo estructural a partir de los fragmentos que lo constituyen. La estructura del complejo TFIIB-TBP-ADN (código PDB: 1C9B) contiene el segmento nucleotídico al cual se unen las proteínas. Esta secuencia se muta para mantener la correspondencia con la hebra total de ADN a modelar. Se elimina la estructural de la proteína TFIIB del complejo. Se generan dos fragmentos de ADN contexto (con estructura doble hebra en conformación B), cuya longitud y secuencia dependerá del tamaño en pares de bases de la región promotora. Estos fragmentos contienen bases que solapan los extremos del ADN en la proteína a conectar. Los fragmentos se ensamblan mediante alineamiento estructural de los extremos solapantes de ADN. Una vez generado el alineamiento se eliminan las bases repetidas para conservar la continuidad del polímero de ADN.

Finalmente, los modelos simplificados se generaron removiendo y renombrando átomos de acuerdo al esquema definido en la Figura 1 del Artículo 1 sección Publicaciones [45]. En el caso del sistema multiescala el procedimiento fue igual al ya descrito para los sistemas simplificados con proteínas, con la diferencia de que se conservo a la proteína TBP y no se modificaron los residuos de la región atomística. A modo de ejemplo, la Figura 1.2.3 muestra los modelos estructurales resultantes para los sistemas S_{VIH}, $S_{VIH+TBP}$ y $S^*_{VIH+TBP}$.

En la simulación de los sistemas se empleó el paquete AMBER11 [23]. Se emplearon los parámetros desarrollados por Dans *et al*. y Darré *et al.* (Artículos 1 y 3 sección Publicaciones, [45,46]) para describir los modelos simplificados. En el sistema multiescala se utilizaron adicionalmente los campos de fuerza parm99SB [71] y parmbsc0 [72] en la región atomística y los parámetros desarrollados por Machado *et al*. (Artículo 5 sección Publicaciones, [66]) para la interfase.

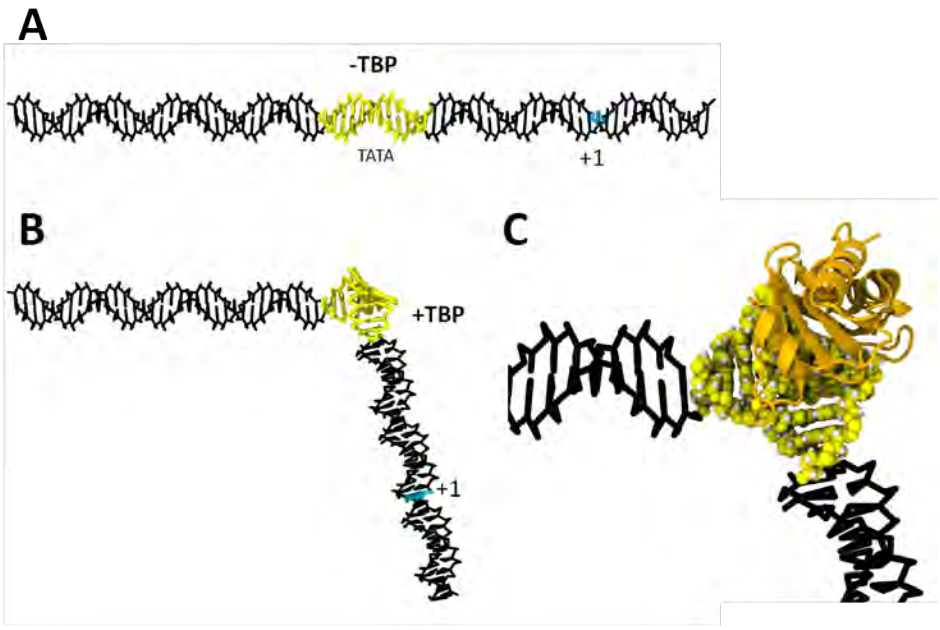

FIGURA 1.2.3 | Modelos estructurales de la región promotora del virus VIH-1. A) Sistema sin TBP (S_{VIH}). En amarillo se representa la región de unión a TBP mientras que el par de bases del sitio +1 se muestra en azul. B) Modelo simplificado de la unión de TBP a la secuencia promotora de VIH-1, sistema $S_{VIH+TBP}$. La proteína se modela de forma implícita considerando solo los cambios estructurales que genera sobre el ADN al unirse. C) Ampliación de la región atomística en el sistema multiescala $S^*_{VIH+TBP}$. Se pueden distinguir los átomos pesados que forman las bases así como los átomos de hidrógeno, esferas amarillas y blancas respectivamente. La proteína TBP, descrita explícitamente a nivel atómico, esta representada como cintas de color naranja. El resto de la secuencia promotora es modelado de forma análoga al sistema $S_{VIH+TBP}$.

La configuración inicial de todos los sistemas fue sujeta a 1500 pasos de minimización. El protocolo de DM para los sistemas simplificados consistió en una etapa de calentamiento desde 0 K a 298 K en 500 ps de simulación, seguida de 10 μs de producción. Se acopló el sistema al termostato de Langevin usando como valor para el coeficiente de fricción 50 ps^{-1}. Se empleó un paso de integración de 20 fs y se recolectaron configuraciones cada 100 ps para el análisis. Los efectos de hidratación y la fuerza iónica se tuvieron en cuenta implícitamente con el modelo de Born Generalizado [38-40] y Debye-Hückel [73]. Se uso una concentración salina de 0.15 M para representar el entorno fisiológico. Se empleó como distancia de corte 18 Å para el cálculo de las interacciones electrostáticas. Para simular la continuidad de la doble hebra, se añadieron restricciones armónicas débiles (3.0 Kcal mol^{-1} $Å^{-2}$) a modo de preservar los enlaces de hidrógeno de Watson-Crick en los pares de bases de los extremos. En los sistemas con TBP, la interacción ADN-proteína se modeló de forma implícita empleando restricciones posicionales de 0.1 Kcal mol^{-1} $Å^{-2}$

sobre la configuración inicial (estructura del cristal) en los pares de base de la región de unión a TBP (ver Figura 1.2.3 B).

En la simulación del sistema multiescala ($S^*_{VIH+TBP}$) el protocolo fue esencialmente igual al ya descrito, con la diferencia de que se empleó el algoritmo LINCS para restringir la distancia de enlace entre átomos pesados e hidrógenos y se utilizo un paso de integración de 2 fs, generándose 320 ns de producción. Se empleó el paquete AMBER11 en su versión para GPU. No se usaron restricciones posicionales sobre la región de unión a TBP ya que la proteína se representa de forma explicita (ver Figura 1.2.3 C).

Para identificar regiones con comportamiento particular en el ADN se evaluaron propiedades en ventanas de 5 pares de base de longitud. La elección de este tamaño de ventana se basa en un criterio funcional. Según observaciones de Alexandrov *et al.* la presencia de 5 pares de bases desapareados consecutivos son suficientes para iniciar la transcripción por la RNA polimerasa II [74]. Desde un punto de vista estructural, este tamaño de ventana también corresponde al ancho de los surcos del ADN en la conformación B. El análisis se realizó desde el extremo 5' al 3' de la secuencia, moviendo la ventana de estudio de a un par de base por vez. Se cuantificó la cantidad de eventos extremos de: desapareamiento de bases (respiración), torsión del esqueleto y apertura o cierre del surco mayor y menor. Los eventos se definieron a partir de medidas estructurales en el ADN, en cada caso se definió un umbral para discretizar las observaciones. Se definió como par de base desapareado cuando la distancia *r* entre los átomo centrales que forman la interacción Watson-Crick es mayor a 4 Å (ver Figura 1.2.4 A). Según se demostró este criterio permite describir correctamente temperaturas de fusión en el ADN [45,63]. Se computó como un evento de respiración en la doble hebra de ADN cuando todos los pares de bases de la ventana de estudio se encontraron desapareados simultáneamente. Por otro lado, las torsiones y longitudes de los surcos pueden ser tanto positivas, negativas o nulas según sean mayores, menores o cercanas al valor medio de la propiedad. En este caso se tomo como criterio computar eventos que cumplan con la condición: $x < (\mu-2\sigma)$ ó $x > (\mu+2\sigma)$, donde x es la propiedad a medir, μ es el promedio de la

misma y σ el desvío estándar. Al restar ambas cantidades se obtiene, en una ventana particular, la dirección neta de la propiedad observada.

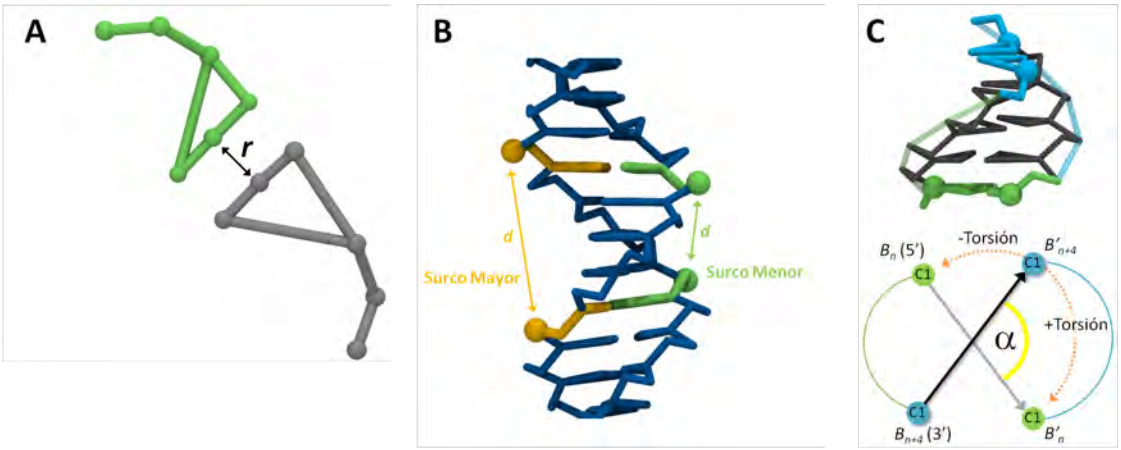

FIGURA 1.2.4 | Propiedades estructurales medidas en el ADN. A) El apareamiento de un par de base se mide usando la distancia *r*, la cual se define a partir de los centroides centrales que forman el par Watson-Crick. B) El tamaño del surco mayor o menor se mide como la distancia *d* entre centroides que representan al grupo fosfato del ADN. Dichos centroides se muestran como esferas en el modelo estructural. Las bases empleadas como referencia determinan cual es el surco a medir. El surco mayor se mide entre una base de la hebra líder y la base de la hebra complementaria que se encuentra a 5 nucleótidos de distancia en dirección 3'. El surco menor se mide de forma análoga al surco mayor pero empleando la hebra retrasada. C) La torsión en la doble hebra de ADN se mide como el giro relativo entre dos pares de bases, representadas en color azul y verde en el modelo estructural. Las esferas corresponden a centroides en la posición C1 de cada base. Las líneas azul y verde representan la continuidad del esqueleto de ADN. El esquema inferior corresponde al modelo estructural visto desde arriba. Para medir el giro relativo de un par de bases respecto a otro se emplea el ángulo α, el cual se forma entre los ejes de cada par de bases. El eje de un par de bases esta definido como un vector que pasa entre los centroides C1 de las bases que lo forman ($B_n \rightarrow B'_n$ o $B_{n+4} \rightarrow B'_{n+4}$, donde *n* hace referencia a la posición en la secuencia). La dirección 5'-3' de la hebra líder se indica entre paréntesis. La disminución del ángulo α genera torsiones positivas (desenrollamiento de la doble hebra) mientras que el aumento implica torsiones negativas.

La apertura o cierre de los surcos se midió a partir de la distancia *d* entre centroides correspondientes a los grupos fosfatos del ADN (Figura 1.2.4 B). Valores de $d > (\mu+2\sigma)$ dan lugar a eventos de apertura, mientras que $d < (\mu-2\sigma)$ son eventos de cierre. En el caso de las torsiones la propiedad a medir es el ángulo α definido entre el vector C1-C1 de un par de base respecto a otro 5 pares de base distante en dirección 3' (Figura 1.2.4 C). Según se puede observar en la Figura 1.2.4 C, valores de $\alpha < (\mu-2\sigma)$ dan lugar a eventos de torsiones positivas mientras que $\alpha > (\mu+2\sigma)$ genera eventos de torsión negativa (enrollamiento de la doble hebra). La Tabla 1.2.1 muestra cuales son los valores de μ y σ característicos de cada propiedad y por ende usados en este estudio. Estos valores corresponden a medidas realizadas a partir de varias secuencias simuladas (ver Artículo 1 sección Publicaciones, [45]).

TABLA 1.2.1 | Distribución de valores para propiedades estructurales medidas en simulaciones de varias secuencias [45]. (a) Ángulo α y (b) distancia *d* medidos según la Figura 1.2.4.

Propiedad	Promedio	Desvío estándar
[a]Torsión de la doble hebra	135°	5°
[b]Tamaño del surco mayor	19 Å	1.5 Å
[b]Tamaño del surco menor	12 Å	1 Å

Para poder comparar entre distintas simulaciones se normalizó el numero de eventos observado al porcentaje de ocurrencia (%Occ). Esto se hizo simplemente considerando el porcentaje que ocupa el número de eventos observados en el total de configuraciones analizadas.

Para el sistema multiescala también se evaluó la raíz de la desviación cuadrática media (*del inglés: root mean square deviation,* RMSD) del ADN en la región de unión a TBP (región atomística) respecto a la estructura cristalográfica 1C9B. Dados dos grupos de átomos distribuidos en el espacio con los mismos elementos *v* y *w*, el RMSD de *v* respecto de *w* se mide como:

$$RMSD(v,w)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2+(v_{iy}-w_{iy})^2+(v_{iz}-w_{iz})^2} \quad (9)$$

Donde n es el número total de átomos en *v* o *w* y los subíndices *x*, *y*, *z* refieren a los componentes de la posición de cada átomo *i* en el sistema cartesiano. Esta medida cuantifica cuan distantes son las conformaciones obtenidas durante la dinámica respecto a la referencia. Para el cálculo se emplearon todos los átomos en el ADN que componen la región de unión a TBP.

Resultados

La unión de TBP a la secuencia TATA esta caracterizada por una serie de cambios estructurales a nivel del ADN. TBP se une al surco menor del ADN, induciendo en este un quiebre cercano a 90° en dirección al surco mayor [58]. A su vez las bases en contacto con la proteína pasan de encontrarse en la conformación B a la A del ADN luego de la interacción. Durante el mecanismo de unión se requiere la presencia de otros factores como TFIIA y TFIIB para estabilizar la interacción de TBP con el ADN [56]. Estas modificaciones estructurales son esenciales para preparar al ADN con el fin de iniciar la transcripción.

Simular un sistema molecular completamente a nivel simplificado es muy conveniente desde el punto de vista computacional. Sin embargo es necesario comprobar que las aproximaciones empleadas son capaces de capturar el fenómeno a estudiar. Si bien anteriormente se demostró la capacidad del modelo para reproducir el comportamiento del ADN (ver Artículos 1 al 4 sección Publicaciones), es necesario comprobar que se puede modelar implícitamente la interacción TBP-ADN sin introducir artefactos. En particular es preciso demostrar que usar restricciones posicionales sobre la región de unión a TBP cuando el ADN se encuentra en la conformación de enlace a la proteína, es suficiente para emular la perturbación sobre el resto del polímero. Esto implica comparar la descripción explícita e implícita de la proteína en la región de unión, lo cual se realizó estudiando los sistemas $S^*_{VIH+TBP}$ y $S_{VIH+TBP}$ respectivamente.

Lo primero que se observó en la simulación explícita del sistema con TBP ($S^*_{VIH+TBP}$) es que el ADN en la región de unión a la proteína se mantiene en una conformación cercana a la presente en el cristal del complejo TBP-ADN. De hecho, los valores de RMSD sobre esta región se apartan en torno a 3.5 Å de la estructura cristalográfica (Figura 1.2.5). Esto implica que la proteína TBP es capaz de restringir de forma importante la dinámica y conformación de la región de unión. Esta relativa rigidez puede ser tenida en cuenta implícitamente restringiendo la movilidad de los residuos que componen la región de unión a TBP a la conformación del complejo proteína-ADN. Queda por investigar si la presencia explícita de la proteína genera efectos que inciden sobre las propiedades a medir. En la Figura 1.2.6 se compara el perfil de respiración de los sistemas $S^*_{VIH+TBP}$ y $S_{VIH+TBP}$. La similitud entre ambos perfiles es notable, observándose la mayor diferencia en los extremos de la secuencia. Esto es debido al uso de restricciones armónicas en el sistema $S_{VIH+TBP}$ para mantener el apareamiento de las bases de los extremos. La ausencia de estas restricciones en el sistema $S^*_{VIH+TBP}$ aumenta la probabilidad de observar eventos de respiración en esas regiones. Sin embargo estas diferencias se restringen a unos pocos pares de bases. Por lo tanto, los resultados presentados validan la simulación implícita de TBP mediante restricciones posicionales en los sistemas simplificados. En lo que resta de la sección se

presentan los resultados comparativos entre sistemas con y sin TBP unida. Estas observaciones se basan en sistemas simplificados y la descripción implícita de la proteína.

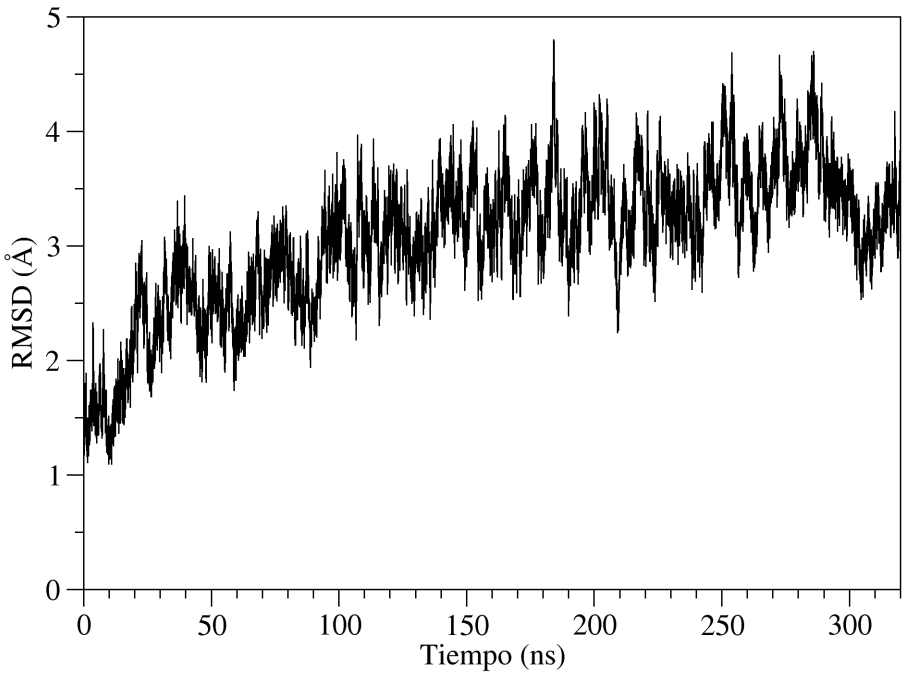

FIGURA 1.2.5 | Valores de RMSD calculados sobre la región de unión a TBP en la simulación del sistema $S^*_{VIH+TBP}$. Se utilizó como referencia la estructura del complejo TFIIB-TBP-ADN (1C9B).

Continuando con el estudiando el promotor del VIH-1, la Figura 1.2.7 A muestra como se ve afectado el perfil de respiración del promotor al unirse TBP a la región TATA. El principal efecto implica un aumento en la probabilidad de apertura de los 8 pares de bases contiguos a la región de unión a TBP en dirección 3' (ventanas comenzando en posiciones -18, -17, -16 y -15 de la secuencia). Claramente este efecto es asimétrico y de corto alcance. Por otro lado el cambio de torsión en el ADN afecta principalmente los 8 pares de bases en dirección 5' de la región de unión a TBP (ventanas comenzando en posiciones -38, -37, -36 y -35 de la secuencia), e implica eventos tendientes a desenrollar el ADN (Figura 1.2.7 B). La propiedad más afectada por la presencia de TBP es la apertura y cierre del surco mayor (Figura 1.2.7 C). Al unirse la proteína se invierte el patrón observado hacia el extremo 3' desde la caja TATA. Este efecto se extiende aproximadamente 25 pares de bases

alcanzando el inicio de transcripción. Mientras que hacia el extremo 5' el efecto se extiende unos 20 pares de bases. En este caso no es claro que se invierta el perfil de apertura y cierre, pero si existe un cambio importante en la población de varios picos del gráfico. En particular aumenta notoriamente la apertura del surco mayor en las bases contiguas a la región de unión a TBP. Por el contrario, no se aprecian efectos notables sobre el surco menor (Figura 1.2.7 D).

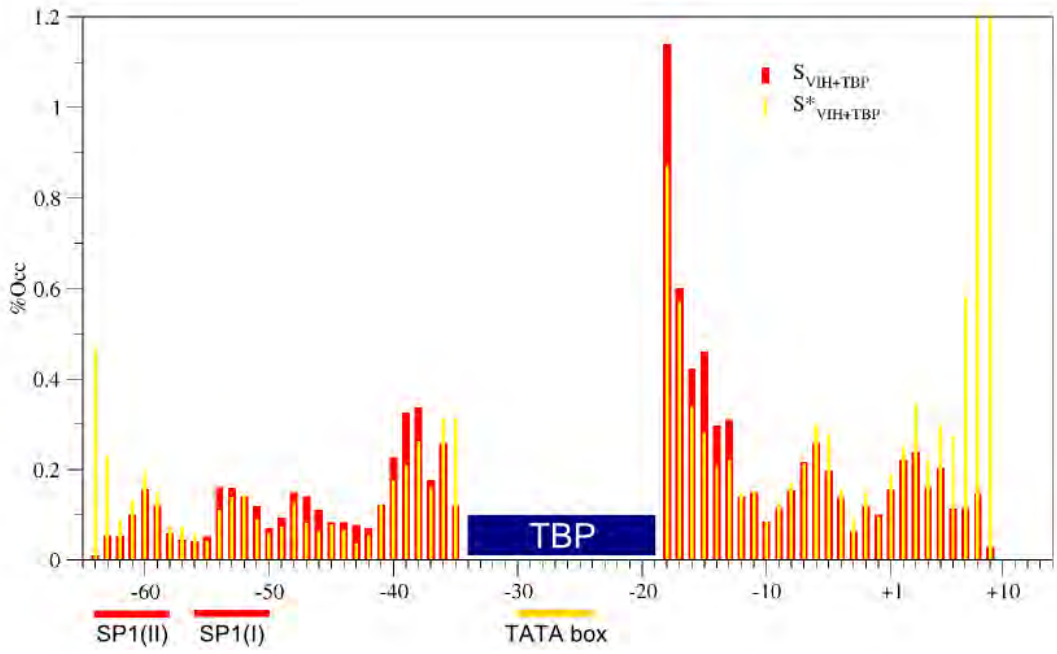

FIGURA 1.2.6 | Porcentaje de ocurrencia (%Occ) de eventos de respiración en los sistemas $S_{VIH+TBP}$ (rojo) y $S^*_{VIH+TBP}$ (amarillo). Cada barra corresponde a la medida obtenida para una ventana de 5 pares de bases, cuya secuencia comienza en el nucleótido bajo la barra y se extiende en dirección 3' (derecha). La región que ocupa TBP se indica con un rectángulo azul en el centro de la figura. Debajo de la secuencia se señalan los distintos elementos reguladores presentes en la región promotora.

En términos generales se puede decir que la dinámica de la región promotora de VIH-1 se ve afectada por la presencia de TBP. La magnitud del efecto varia según la propiedad observada. Mientras las modificaciones en los perfiles de respiración y torsión son próximas a la región donde se une la proteína, los cambios en el surco mayor pueden extenderse a 20 pares de bases de distancia de la proteína. Es importante notar que factores de transcripción como TFIIB unen secuencias aledañas a TBP [75]. En particular, según se puede observar de la estructura del complejo TFIIB-TBP-TATA en el contexto del promotor, el factor TFIIB hace contacto con el surco mayor del ADN hacia la región 5' de TBP y con el surco menor hacia el extremo 3' de la

misma (Figura 1.2.8). De forma coincidente, las ventanas de ADN próximas al lugar de contacto de TFIIB (posiciones -38 a -35 en la secuencia) muestran mayor apertura del surco mayor cuando se encuentra unida TBP (Figura 1.2.7 C). Este efecto se ve acompañado de un aumento en eventos de torsión positiva en el ADN (Figura 1.2.7 B). A su vez, las ventanas en las posiciones -16 a -13 del surco menor, también modifican (aunque en menor medida) su comportamiento (Figuras 1.2.7 D). Algunas de ellas tienden a estar más comprimidas (ventana en posiciones -16 y -15), lo cual se ha visto es una señal para proteínas que reconocen el surco menor [16]. Ensayos de metilación demostraron que la posición -16 es un sitio de unión específica en VIH [76]. Sin embargo otras ventanas parecen mostrar un estado más abierto. Estos cambios en la dinámica de los surcos podría favorecer el enlace de TFIIB al ADN durante la interacción con TBP.

Para comprobar si todas estas observaciones son específicas al promotor del VIH-1 o pueden ser generalizadas a otros casos se estudió el promotor SCP1. Este promotor constituye un caso completamente opuesto al VIH-1. No solo no depende de potenciadores para la transcripción, sino que posee una tasa de transcripción 8 veces mayor al promotor de *Cytomegalovirus* [68].

Al contrario de lo observado para el promotor de VIH-1, el perfil de respiración en SCP1 se ve menos alterado en la unión con TBP (Figura 1.2.9 A). Las dos regiones que presentan mayores alteraciones en función del enlace a TBP, se encuentran delimitadas por el elemento TATA y el origen de transcripción, y por este último y el elemento MTE. Sin embargo, salvo por la ventana que comienza en la posición +15 de SCP1, el resto presenta valores de respiración muy por debajo de los observados para las ventanas de -38 a -35 del promotor de VIH.

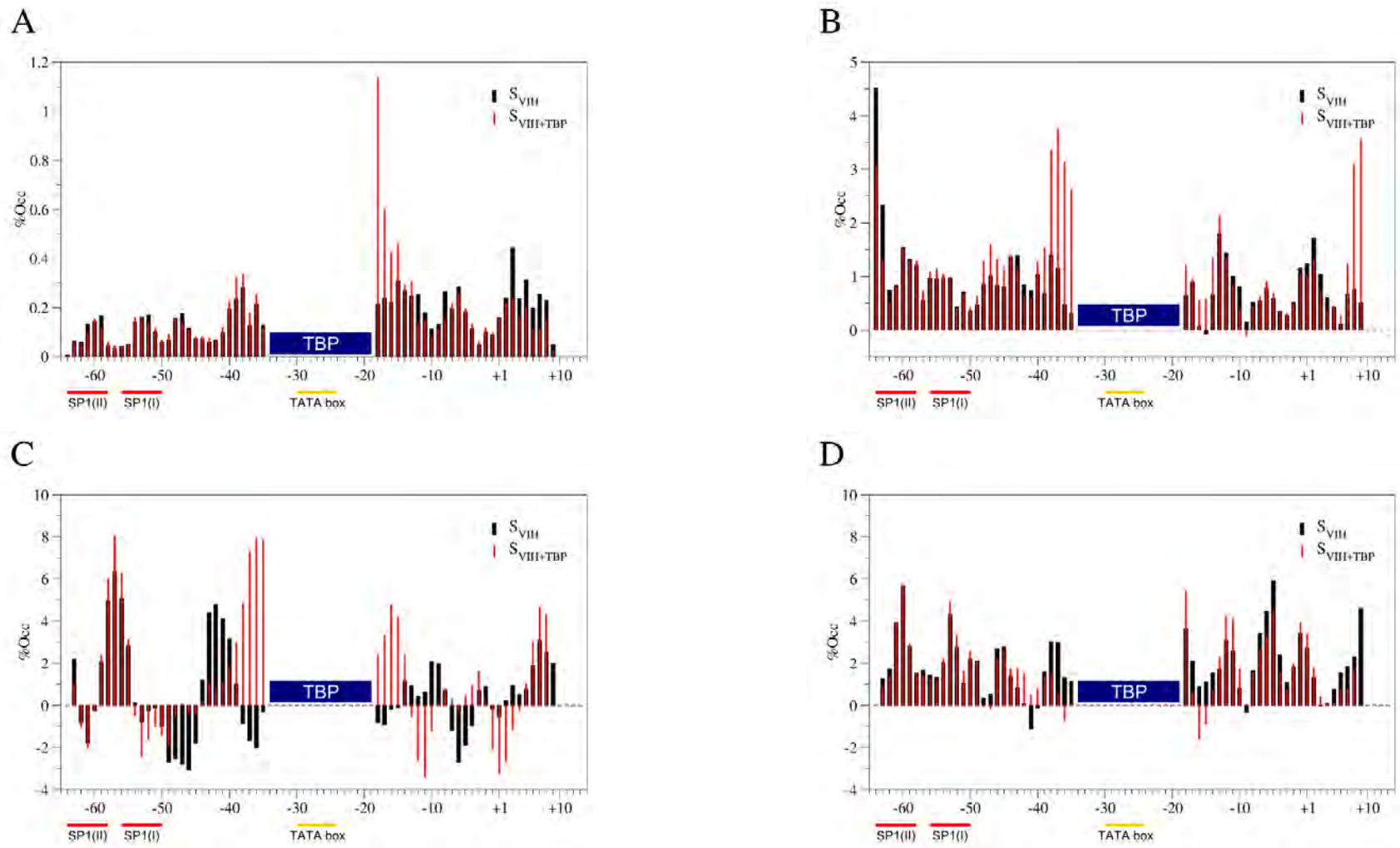

FIGURA 1.2.7 | Porcentaje de ocurrencia (%Occ) de eventos en diferentes propiedades medidas para los sistemas S_{VIH} (negro) y $S_{VIH+TBP}$ (rojo). (A) Perfiles de respiración. (B) Eventos de torsión. Los valores positivos indican torsiones positivas (desenrollamiento de la doble hebra), mientras valores negativos corresponden a torsiones negativas. (C, D) Apertura o cierre de surco mayor y menor respectivamente. En ambos casos valores positivos indican apertura del surco (aumento del tamaño), mientras que valores negativos son eventos de cierre. Cada barra corresponde a la medida obtenida para una ventana de 5 pares de bases, cuya secuencia comienza en el nucleótido bajo la barra y se extiende en dirección 3' (derecha). La región que ocupa TBP se muestra en azul. Debajo de la secuencia se señalan los distintos elementos reguladores presentes en la región promotora.

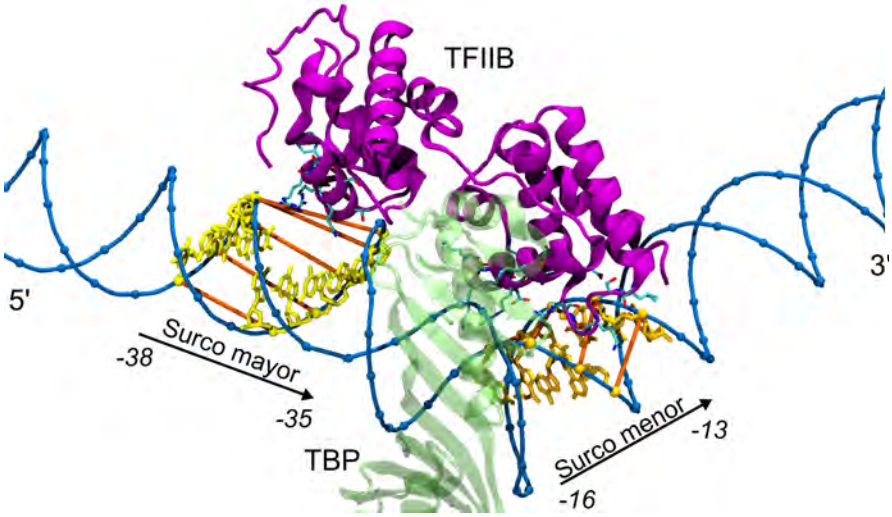

FIGURA 1.2.8 | Complejo TFIIB-TBP-TATA en el contexto de la secuencia promotora de VIH-1. El esqueleto fosfato del ADN esta representado mediante tubos y esferas azules. El factor TFIIB (purpura) interacciona con el surco mayor (amarillo) hacia el extremo 5' de TBP (verde) y con el surco menor (naranja) hacia del extremo 3' de la secuencia. Se muestran las bases que forman las ventanas próximas a los sitios de unión de TFIIB. Las barras naranjas que conectan grupos fosfato ente bases, indican como se define cada ventana. El rango de ventanas para cada región de interacción se indica con una flecha, los valores corresponden a la posición inicial de las ventanas tomando como referencia la hebra líder (ver Figura 1.2.7). El resto de los pares de bases se omitieron por simplicidad.

Los cambios más notables generados por la unión de TBP al promotor se pueden observar en la torsión del ADN y en el tamaño del surco mayor (Figura 1.2.9 B y C). Ambas propiedades estructurales muestran como las ventanas -16 a -5 y +9 a +23 aumentan su accesibilidad en términos de apertura del surco mayor y torsión positiva del ADN. En ambos casos las modificaciones en los perfiles se extienden a más de 20 pares de bases del sitio de unión a TBP. De ambas propiedades, el cambio más notable se ve en el surco mayor pasando de un perfil global cerrado a abierto. Es interesante notar que según trabajos recientes, estas regiones corresponden a sitios de anclaje para TFIID [68]. Una vez más el tamaño del surco menor parece ser una propiedad menos sensible a la unión de TBP comparado con el surco mayor (Figura 1.2.9 D). En este punto se debe destacar el claro aumento en estados comprimidos del surco menor sobre la posición +17. Este efecto se correlaciona con el gran aumento de eventos de apertura del surco mayor en esa ventana (Figura 1.2.9 C).

En ninguno de los casos se aprecian efectos hacia el extremo 5' desde la región de unión a TBP. Tampoco se evidencian patrones análogos al promotor de VIH-1 en las regiones de interacción con TFIIB.

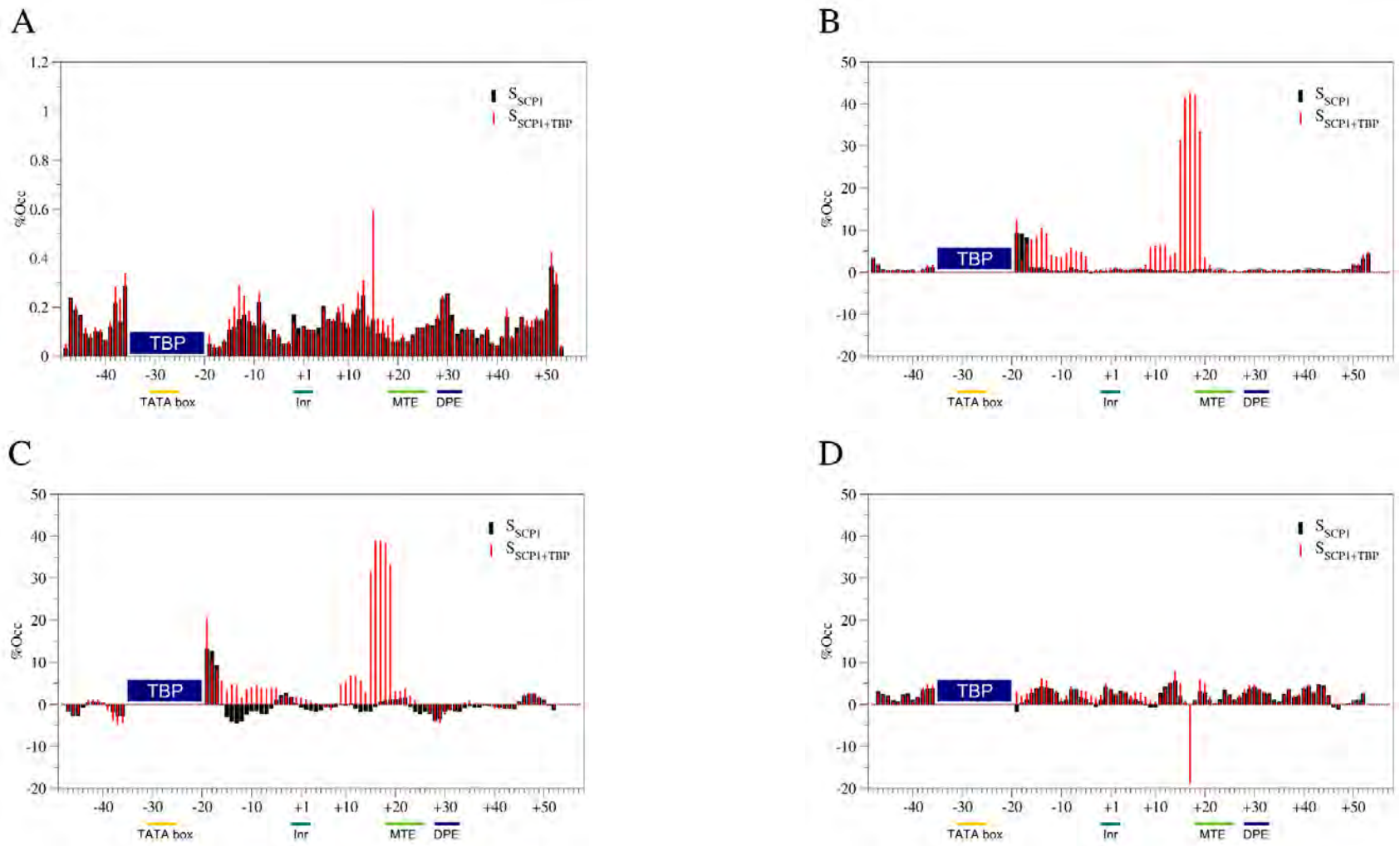

FIGURA 1.2.9 | Porcentaje de ocurrencia (%Occ) de eventos en diferentes propiedades medidas para los sistemas S_{SCP1} (negro) y $S_{SCP1+TBP}$ (rojo). (A) Perfiles de respiración. (B) Eventos de torsión. Los valores positivos indican torsiones positivas (desenrollamiento de la doble hebra), mientras valores negativos corresponden a torsiones negativas. (C, D) Apertura o cierre de surco mayor y menor respectivamente. En ambos casos valores positivos indican apertura del surco (aumento del tamaño), mientras que valores negativos son eventos de cierre. Cada barra corresponde a la medida obtenida para una ventana de 5 pares de bases, cuya secuencia comienza en el nucleótido bajo la barra y se extiende en dirección 3' (derecha). La región que ocupa TBP se muestra en azul. Debajo de la secuencia se señalan los distintos elementos reguladores presentes en la región promotora.

Promotores tan diversos como VIH-1 y SCP1 pueden dar lugar a comportamientos diferentes al estar diseñados para responder a vías distintas de activación. Por ello, para evaluar el efecto de cambios pequeños en la secuencia se estudió el promotor m1SCP1. Este contiene cuatro mutaciones en la secuencia de SCP1 que bajan el rendimiento de la transcripción sin impedir la unión de factores de transcripción TBP, TFIIA, TFIIB, TFIID entre otros [69]. El resultado para m1SCP1 fue concluyente, los perfiles de todas las propiedades prácticamente no cambiaron al unirse TBP (Figura 1.2.10). Tanto el incremento de la torsión positiva como apertura del surco mayor pasan a tener un impacto local que no se extiende de la ventana en la posición -15 de la secuencia. Esto sugiere que las mutaciones reducen la capacidad del ADN para responder a la unión de la proteína.

Conclusiones

En este contexto, es importante destacar la magnitud que representa cada tipo de propiedad medida sobre el ADN. Mientras los eventos de respiración apenas alcanzan el 0.1% del total de la DM, correspondiente a una frecuencia de $1x10^{-3}$ observaciones, los efectos en torsiones y surcos pueden superar el 5-10% y en algunos casos llegar al 40% de ocupación. Recientemente, Alexandrov *et al.* utilizando un modelo matemático unidimensional no lineal, lograron correlacionar de forma exitosa los niveles de respiración en el inicio de transcripción con la tasa de transcripción [74]. En su modelo los eventos medidos poseían una frecuencia menor a $1x10^{-4}$ sin embargo, estos parecen ser suficientes para cumplir un rol biológico. Si bien en el modelo propuesto en este trabajo se obtuvieron conclusiones cualitativamente equivalentes, en nuestro caso la sensibilidad para detectar cambios en la dinámica del ADN es 1 a 3 ordenes mayor (en el caso de torsiones y tamaño de surcos). Es de destacar que el modelo propuesto por Alexandrov *et al.* es de índole puramente matemática, mientras que el desarrollado en este trabajo permite obtener además propiedades estructurales que abarcan la mayoría de los fenómenos presentes en el ADN. Esto permite proponer a la metodología empleada como una herramienta muy útil para entender la fenomenología detrás de cambios en el ADN, así como su evolución espacial y temporal.

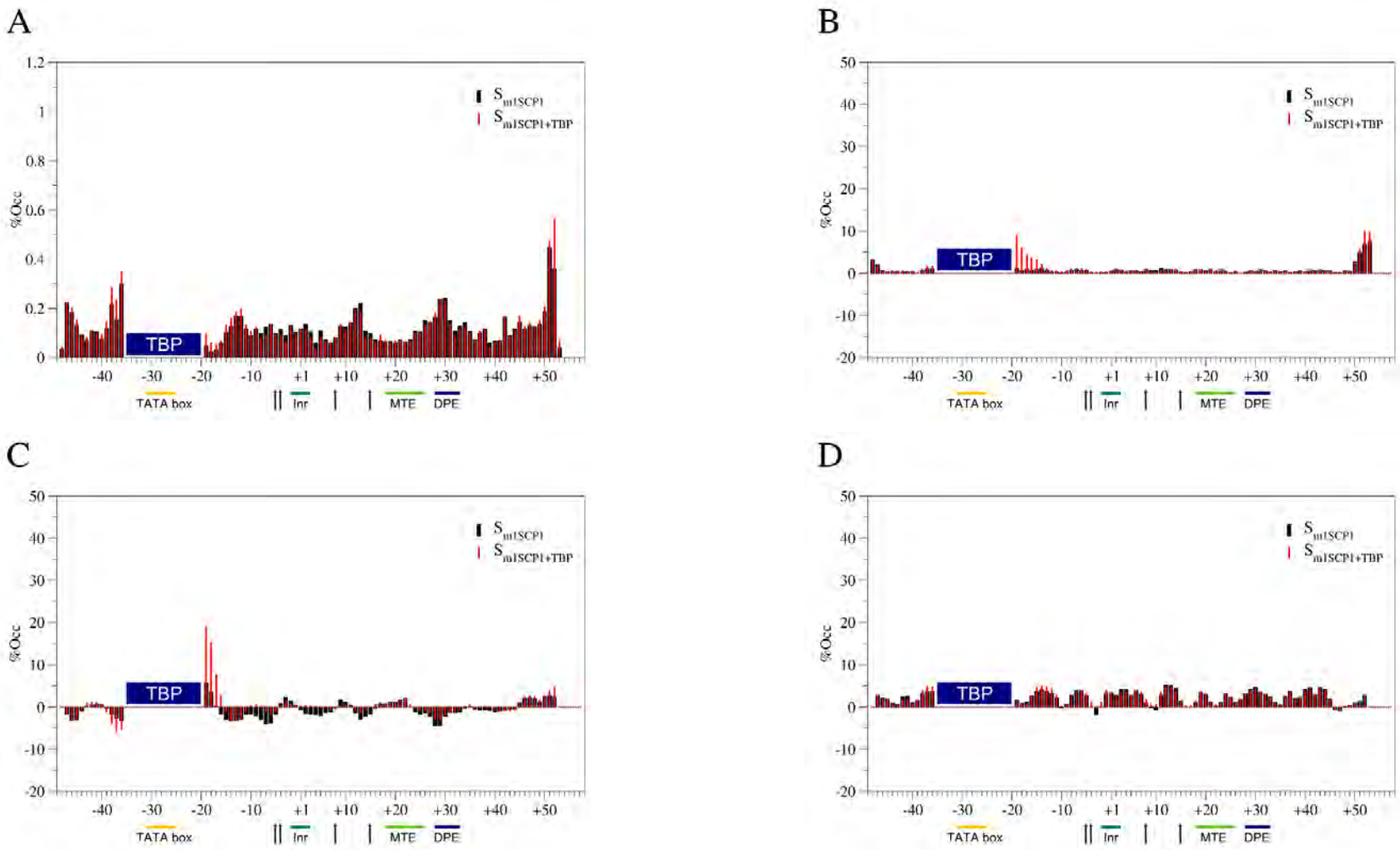

FIGURA 1.2.10 | Porcentaje de ocupación (%Occ) de eventos en diferentes propiedades medidas para los sistemas S_{m1SCP1} (negro) y $S_{m1SCP1+TBP}$ (rojo). (A) Perfiles de respiración. (B) Eventos de torsión. Los valores positivos indican torsiones positivas (desenrollamiento de la doble hebra), mientras valores negativos corresponden a torsiones negativas. (C, D) Apertura o cierre de surco mayor y menor respectivamente. En ambos casos valores positivos indican apertura del surco (aumento del tamaño), mientras que valores negativos son eventos de cierre. Cada barra corresponde a la medida obtenida para una ventana de 5 pares de bases, cuya secuencia comienza en el nucleótido bajo la barra y se extiende en dirección 3' (derecha). La región que ocupa TBP se muestra en azul. Debajo de la secuencia se señalan los distintos elementos reguladores presentes en la región promotora, las flechas indican los sitios mutados respecto al promotor SCP1.

La unión de TBP al promotor generó efectos visibles que en algunos casos alcanzaron más de 20 pares de base de distancia del sitio de enlace a la proteína. Este efecto tiene una clara componente secuencia dependiente que se observó a diferentes niveles. Promotores tan disimiles como el VIH-1 y SCP1 muestran comportamientos claramente diferentes al unir TBP. Este comportamiento puede deberse a las diferentes vías de activación que emplea cada promotor. Por un lado el VIH-1 requiere de potenciadores como NF-kb, y factores de transcripción como PTEFb (vía independiente de TFIID, [77]). Mientras que SCP1 es un promotor artificial que no requiere de potenciadores y es dependiente de la unión a TFIID para la transcripción. Por otro lado y de forma notable, unas pocas mutaciones pueden generar grandes cambios a nivel de las propiedades estructurales del ADN, tal como se observo entre SCP1 y m1SCP1. En este sentido, la presencia de polimorfismos tanto en el VIH como en otras secuencias virales podrían tener gran impacto a nivel del ADN.

Por todo lo expuesto, se puede concluir que el ADN no es una plataforma pasiva a la unión de proteínas. La interacción ADN-proteína puede alterar dramáticamente la dinámica de patrones estructurales sobre secuencias distantes al sitio de unión. Nuestro modelo propone además que estas alteraciones estructurales tienen una existencia temporal putativa en el orden de los multi-microsegundos, es decir escalas temporales biológicamente relevantes. Estos cambios podrían constituir por tanto una vía complementaria para la transmisión de señales entre elementos de regulación, por medio de la cual no se afectaría la selectividad pero si la accesibilidad de las proteínas a los sitios de enlace.

CAPÍTULO 2: Mecanismos asociados a la represión y estado de latencia del VIH-1

El estado de represión del genoma proviral esta caracterizado por la presencia de proteínas celulares que favorecen el estado compacto de la cromatina [78]. Señales epigenéticas colaboran a regular este proceso. Estas marcas epigenéticas son modificaciones postraduccionales en las colas de las histonas que pueden extenderse a cientos de nucleosomas y cuya transducción por proteínas impacta en la accesibilidad de miles de pares de bases del ADN [79,80]. En particular, es ampliamente conocido que la metilación promueve el estado represivo de la cromatina mientras que la acetilación estimula la transcripción génica [81]. En este juego, proteínas como la HP1 reconocen señales de metilación y actúan como adaptadores moleculares. Estas facilitan el reclutamiento de proteínas efectoras como metiltransferasas y deacetilasas, las cuales a su vez promueven y mantienen el silenciamiento [82]. El papel que desempeña HP1 hace que sea un moderador clave en la vía de comunicación que conduce a la latencia del virus. Por lo tanto entender como interacciona con otros actores es importante para dilucidar los mecanismos de regulación. En particular este estudio se centró en entender los determinantes isoforma específicos de HP1 en la interacción con H3.

Además de interacciones proteína-proteína, durante el estado de latencia también existe la unión de factores de transcripción a elementos reguladores en regiones accesibles del LTR. La combinación particular de estos factores de transcripción junto a las proteínas que reclutan también contribuye a determinar el estado de expresión del genoma viral [83]. Tener una noción sobre la distribución espacial de los distintos actores que participan en la regulación puede generar una nueva visión sobre las vías de comunicación molecular que dan lugar a la represión del genoma viral.

Las principales preguntas que busca responder este capítulo son: ¿Existen factores estructurales isoforma específicos determinantes en la unión de HP1 a H3? ¿Como es la estructura macromolecular del estado provirus de VIH-1?

Caracterización de determinantes estructurales en la interacción isoforma específica de HP1 con la histona 3

HP1 es el nombre de una familia de proteínas nucleares no histónicas ampliamente conservadas en la escala evolutiva, presentes desde levaduras hasta mamíferos [84]. En humanos existen tres isoformas, alfa (HP1a), beta (HP1b) y gama (HP1g), que difieren en su ubicación en el núcleo celular [85]. Mientras HP1a y HP1b se concentran en la heterocromatina pericéntrica, HP1g también se ubica en sitios de eucromatina [86,87]. Esta proteína posee una estructura modular compuesta por dos dominios homólogos unidos por un segmento no estructurado (ver Figura 2.1.1). Los dominios llamados Chromo (del inglés: *CHRomatin Organization MOdifier*) y Chromo Sombra (en referencia a su homología con el primero) se encuentran ubicados en los extremos amino- y carboxilo-terminal, respectivamente. El dominio Chromo funciona a modo de ancla ligando específicamente la Lys9 de la histona 3 en estado di- o tri-metilado [87]. Mientras tanto, el dominio Chromo Sombra media la homodimerización. Al homodimerizar se genera un sitio de enlace capaz de ligar de forma específica secuencias extendidas con motivo canónico Pro-Xxx-Val-Xxx-Leu en segmentos no estructurados de otras proteínas [88]. Recientemente se ha descrito que a altas concentraciones (mayores a 1 uM) la HP1 es capaz de oligomerizar. Este fenómeno esta mediado por interacciones entre los dominio Chromo de la proteína, efecto distinto a la homodimerización [89]. La oligomerización de HP1 permitiría establecer una red de interacciones entre nucleosomas cercanos que favorece la formación de la cromatina.

La estructura modular de HP1 permite que funcione como un "*adaptador molecular*", reclutando diferentes proteínas a distintas regiones de la cromatina. Por esta razón HP1 esta involucrada en un gran número de procesos biológicos que van desde la organización de la cromatina y la regulación de la expresión génica [90] hasta la reparación de daños en el ADN [91,92], el desarrollo tumoral [93] y la transcripción de retrovirus [94]. Recientemente, en modelos celulares de VIH-1 reprimido constitutivamente, se ha reportado la participación de las isoformas HP1b y HP1g en el mantenimiento del silenciamiento y en el pasaje a la activación de la transcripción del virus [18,19]. Dicho pasaje estaría mediado por un cambio de ocupación entre ambas

isoformas en respuesta a modificaciones postraduccionales en la histona 3 (H3). Particularmente la fosforilación en la Ser10 de H3 sería un factor determinante para disociar la interacción de HP1b con la Lys9 trimetilada de H3 y darle paso a HP1g. La alta identidad de secuencia existente entre HP1b y HP1g en el dominio de enlace a Lys9 (mayor al 80%) hace pensar que de existir factores isoforma específico, estos deberían encontrarse en residuos distintos al sitio de unión.

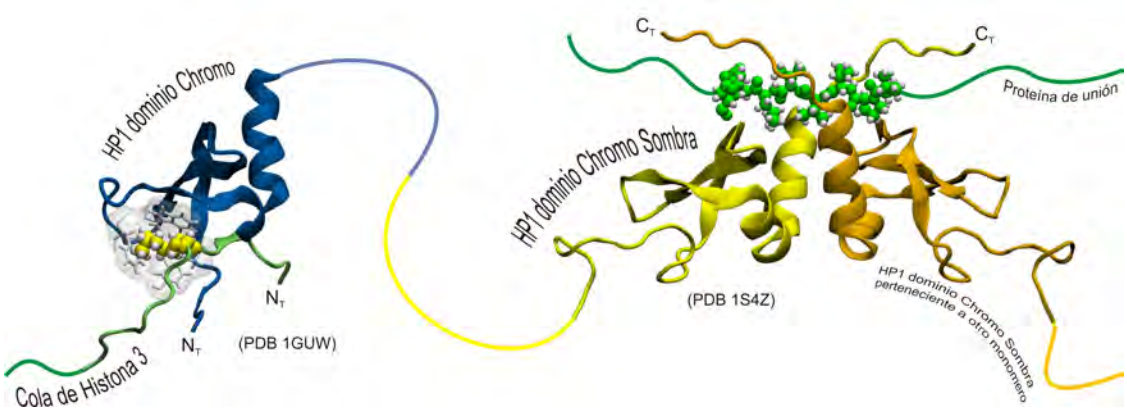

FIGURA 2.1.1 | Estructura de HP1. El dominio Chomo (azul) posee un bolsillo aromático (residuos representados con varillas y superficie gris) capaz de estabilizar una interacción de tipo catión pi. De esta forma la HP1 reconoce la Lys9 de H3 (verde) en estado di o tri-metilada (esferas amarillas y blancas). Una región de aprox. 30 residuos que se presupone poco estructurada conecta el domino Chomo con el Chromo Sombra (amarillo). El dominio Chromo Sombra posee gran afinidad para dimerizar con el dominio análogo de otra HP1 (naranja). En parte de la superficie de dimerización se forma un sitio de unión a otras proteínas como ser metiltransferasas.

Con el objetivo de dilucidar determinantes estructurales que puedan mediar una respuesta isoforma específica a señales epigenéticas, se definió estudiar por DM la interacción del dominio Chromo de HP1b y HP1g con la cola N terminal de H3 (ver Artículo 6 sección Publicaciones, [95]). Se compararon dos escenarios posibles para H3. Uno implicó la trimetilación de la Lys9 (H3K9Me3), mientras que al otros se sumo la fosforilación en la Ser10 (H3K9Me3S10p). La estructura 1GUW [96] se tomo como modelo para construir los sistemas ya que incluye residuos distantes al sitio de reconocimiento de la Lys9 tanto en HP1 como H3.

En las simulaciones realizadas con el sistema H3K9Me3 se lograron observar las interacciones descritas en trabajos experimentales para estas proteínas [96,97] (ver Tabla 1 del Artículo 6 sección Publicaciones, [95]). Ambas isoformas de HP1 mostraron los mismos contactos con los residuos de H3 que conforman la secuencia de unión y dan la especificidad a la interacción

(Gln5-Thr6-Ala7-Arg8-Lys9 [98]). Esto es esperable por la conservación de los residuos que median dicha interacción en ambas isoformas. Sin embargo, la mayor diferencia se dio en la región no estructurada próxima al N terminal de HP1. En particular, se demostró la importancia de los residuos Glu16 y Glu18 de HP1b en la interacción con la Arg8 y la Lys14 de H3. Ambos glutamatos contribuyen a formar un “bolsillo ácido” que media la estabilización de residuos básicos en H3. Esto conlleva a un aumento en la superficie de contacto en la región Lys14. Las mutaciones Glu16Ala y Glu18Pro en HP1b recuperan los patrones observados para HP1g en cuanto a la unión a H3. Tal observación indica la existencia de determinantes estructurales en el extremo N terminal de HP1 para la unión isoforma específica con H3. Recientemente también se ha descrito la importancia del extremo N terminal de HP1 en la interacción de la isoforma HP1a con H3, lo que refuerza la hipótesis propuesta en esta tesis [99]. Estos resultados llevan a proponer un modelo en el cual los residuos 5 al 10 de H3 interaccionan de forma específica y estable con el dominio estructurado de HP1, mientras que la interacción isoforma específica se da entre el extremo N terminal no estructurado de HP1 y entorno de la Lys14 de H3.

Respecto al efecto de la modificación H3K9Me3S10p ambas isoformas se muestran sensibles, lo que se demuestra en la disminución de la energía de enlace, la perdida enlaces de hidrógeno y puentes salinos, y en la reducción de la superficie de contacto. Sin embargo, no fue posible indicar cual isoforma es más susceptible a la fosforilación de la Ser10. En este punto es válido aclarar que los efectos isoforma específicos pueden estar mediados por variados mecanismos, ejemplo modificaciones postraduccionales o interacciones en otras regiones de la proteína además de las estudiadas en esta tesis. A modo de ejemplo, recientemente se ha visto para HP1g que tanto la fosforilación en la Ser83 de su secuencia como la capacidad de esta isoforma para interaccionar con la RNA polimerasa II son factores fundamentales para controlar la transcripción de algunos genes [96,100]. Si bien los determinantes encontrados en este trabajo de tesis pueden no estar implicados en todos los eventos isoforma específicos, muestran que existe un ajuste fino en los mecanismos de regulación por HP1.

Ensamblado molecular del provirus de VIH-1 en estado de latencia

El cambio en la actividad transcripcional del estadio provirus del VIH es el factor clave para entender la persistencia que tiene esta enfermedad [9]. Cuando un provirus pasa a ser transcripcionalmente inactivo, este se convierte en reservorio de nuevas progenies. Distintos mecanismos operan para reprimir la expresión del genoma proviral e inducir un estado de latencia [20]. En linfocitos la regulación génica esta medida principalmente por cambios en el estado de compactación de la cromatina [101]. Puesto que el provirus también esta inmerso en ese contexto puede ser objeto de este tipo de regulación. En el estado latente del provirus se observó una organización particular de la cromatina en torno a la región LTR del VIH-1. En dicha región se encuentran tres nucleosomas ocupando posiciones bien definidas [102]. Cubriendo el inicio del genoma se encuentra el nucleosoma 0 (Nuc0), este abarca los nucleótidos -414 al -254 de la secuencia. Próximo al sitio de iniciación de la transcripción (nucleotidos -2 a +145) se posiciona el nucleosoma 1 (Nuc1). El tercer nuclosoma (Nuc2) se encuentra a 124 pares de base corriente abajo del Nuc1. Estos nucleosomas cumplen un rol muy importante en la expresión del virus pues se localizan en la zona del promotor. Sus efectos no solo se limitan a ocultar o exponer regiones de unión al ADN reconocidas por factores de transcripción, sino que *per se* pueden representar una barrera física para la RNA polimerasa II (RNAP II) y otro componentes de maquinaria de transcripción. Por otro lado, el resto del genoma de aprox. 9Kb de tamaño se encuentra encapsulado dentro de una estructura de cromatina [103]. Diversas proteínas como HP1, metiltransferasas (ej. SUV39H1) y deacetilasas de histonas (HDAC1, HDAC2 y HDAC3), colaboran para mantener este estado de represión. Factores de represión transcripcional como Yin y Yan 1 (YY-1), el factor SV40 tardío (LSF), el corepresor COUP TF (CTIP2), el dimero p50/p50 y SP1 median interacciones ADN-proteína y proteína-proteína involucradas en el reclutamiento de los factores previamente descriptos. Durante la activación de la transcripción aumenta el estado de acetilación de los nucleosomas lo que conlleva al remodelado de la cromatina [8]. Esta etapa esta acompañada por el reclutamiento de acetiltransferasas como p300, CBP, PCAF, etc., por medio de factores como USF, LEF1, el heterodimero p50/p65 y NFAT. Para muchas de

las proteínas mencionadas se conoce la región que ligan en la secuencia del genoma viral [50].

El desarrollo de varias técnicas de biología molecular y estructural junto a la bioinformática ha contribuido a aumentar nuestra comprensión a nivel del genoma, proteoma e interactoma de distintos organismos. Hoy en día la cantidad de estructuras depositadas en el *Protein Data Bank* (PDB) supera las $8x10^4$ y el número de secuencias proteicas en la base de datos *Uniprot* excede las $2x10^7$. Incluso en los casos donde se carece de conocimientos detallados sobre las proteínas o los complejos que forma, es posible emplear herramientas de modelado por homología, dinámica molecular y docking para inferir su ensamblado [104,105]. Por lo tanto, estamos en un punto en el cual conocemos o podemos acceder a varios detalles de los fragmentos que componen los sistemas. Esta posibilidad está permitiendo tender un puente entre la visión atomística de la biología estructural y la perspectiva macroscópica de la biología molecular. Para ello, novedosas aproximaciones se han desarrollado con el fin de integrar datos de variada procedencia en modelos que reflejan la distribución espacial de los componentes [106-108].

En este contexto, lograr una visión estructural de los ensamblados macromoleculares que participan en diversos fenómenos del VIH-1 puede contribuir a comprender mejor la biología de esta enfermedad. Como objetivo se planteó generar un modelo estructural para el genoma proviral de VIH-1 en estado de latencia. Esto implicó hacer una revisión bibliográfica profunda sobre los aspectos estructurales y bioquímicos que se conocen tanto a nivel del virus como de la regulación eucariota.

Metodología

El modelo estructural consistió en el arreglo espacial de la secuencia nucleotídica de VIH-1 y las proteínas unidas en distintas posiciones de la misma. Teniendo en cuenta que en la estructura de los complejos proteicos empleados esta presente el segmento de ADN al cual se unen, es posible usar este ADN como plataforma para guiar el ensamblado macromolecular. La forma más simple de hacerlo es generar una continuidad entre todos los fragmentos de ADN de los complejos usando como vinculo la secuencia a modelar. Con esta aproximación se puede incluir la deformación estructural del

ADN en la región de unión de las proteínas. La secuencia de ADN en los complejos proteicos se mutó de acuerdo a la secuencia total a modelar. Los segmentos de ADN que conectan complejos proteicos se modelaron como fragmentos doble hebra en la conformación B canónica. Estos últimos se construyeron con la utilidad NAB de AMBER11 [23], considerando la secuencia de la región que abarcan. El largo de dichos segmentos incluye 5 pares de bases solapantes en secuencia con el fragmento 3' anterior y 5' posterior a conectar. El ensamblado de todos los fragmentos se realizó mediante alineamiento estructural de los extremos del ADN empleando como referencia los átomos de fosfato en las bases solapantes (ver Figura 2.2.1). Para preservar la continuidad de la doble hebra, se removieron los pares de base repetidos en la región del alineamiento estructural.

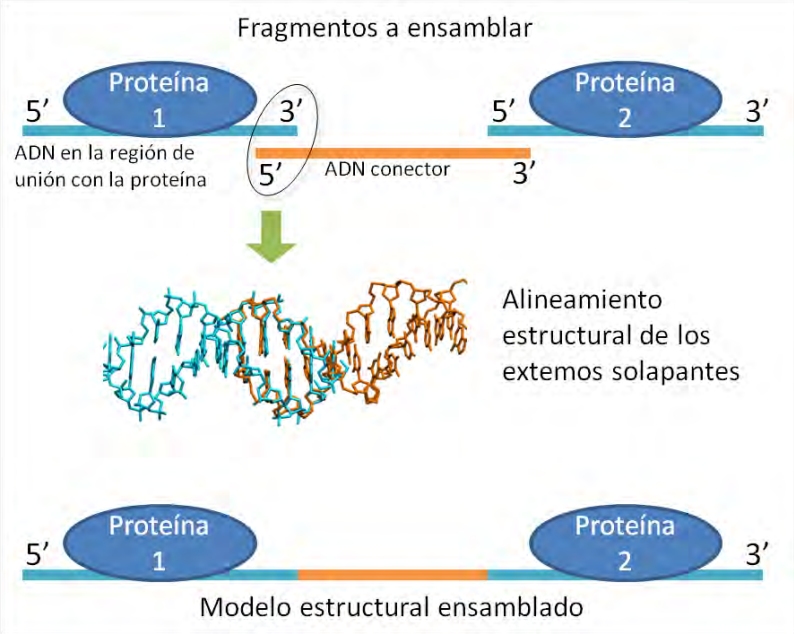

FIGURA 2.2.1 | Ensamblado de complejos macromoleculares sobre el ADN. Se muestra esquemáticamente el procedimiento por el cual se genera un modelo estructural a partir de los fragmentos que lo constituyen. Las estructuras de las proteínas 1 y 2 deben contener los segmentos de ADN al cual se unen. La secuencia de estos segmentos se muta para mantener la correspondencia con la hebra total de ADNmodelar. Se genera un fragmento de ADN para conectar ambos complejos proteicos, cuya longitud dependerá del espaciamiento en secuencia entre las proteínas. Estos fragmentos contienen bases que solapan los extremos del ADN en las proteínas a conectar. Los fragmentos se ensamblan mediante alineamiento estructural de los extremos solapantes en el ADN. A modo de ejemplo se muestra el alineamiento del extremo 3' en la proteína 1 con el extremo 5' del ADN conector. Una vez generado el alineamiento se eliminan las bases repetidas para conservar la continuidad del polímero de ADN.

En el modelo se empleó la secuencia del genoma de VIH-1 (código GenBank: K03455) que consiste en aprox. 10 kilobases. Esta secuencia fue extendida con 2900 residuos de adenina hacia el extremo 5' y 3050 hacia el extremo 3', para representar el contexto estructural de integración en el cual se encuentra el provirus. El resultado final es un genoma total de aprox. 16 kilobases. En la Tabla 2.2.1 se definen las posiciones que ocupa cada complejo proteico en la secuencia del genoma modelado. El complejo de iniciación de la transcripción formado por la RNA polimerasa II (RNAP II), TBP y otros factores se construyo según Bushnell *et al.* [109]. Esto implicó hacer un alineamiento estructural de los segmentos de TFIIB presente en las estructuras reportadas en el PDB con los códigos 1C9B y 3K7A. En el caso de SP1, al no existir estructuras resueltas ligadas a un segmento de ADN, se empleó como modelo el dominio de dedos de zinc del factor TFIIIA (PDB: 1TF3). Si bien la similitud aminoacidica de SP1 y TFIIIA en la región de unión al ADN es de 45%, no se realizo el modelado por homología, sino que se uso directamente esa estructura de TFIIIA como equivalente de SP1. Esta aproximación no implica una perdida demasiado importante de detalle en relación a la resolución total del modelo macromolecular final. En el modelado de la cromatina se empleó una distancia entre nucleosomas de 35 pares de bases, que corresponde al valor medio observado para linfocitos [110]. Para generar una curvatura en la región codificante del VIH se repitió el siguiente patrón de longitudes entre nucleosomas: 35x13, 36x1, 35x4, 36x1, 35x9, 34x1. Donde el numero después del multiplicador 'x' indica la cantidad de veces que se repitió un segmento conector de cierta longitud en pares de bases.

Para estudiar el impacto de la longitud del ADN inter-nucleosoma en la estructura de la cromatina se ensamblaron arreglos de 12 nucleosomas empleando la metodología previamente descripta. En cada arreglo se uso una única longitud inter-nucleosoma para todo segmento de ADN espaciador. Se exploró un barrido de longitudes inter-nucleosoma desde 15 a 60 pares de base tomadas cada 5 bases. El ADN total consistió en una secuencia artificial de poli-adenina suficientemente larga para permitir la estructuración de los 12 nucleosomas en la fibra de cromatina.

TABLA 2.2.1 | Ubicación en la secuencia de los complejos proteicos. Las posiciones se expresan relativas al sitio de iniciación de la transcripción de VIH-1. Las posiciones fueron extraídas de las referencias [a] [102], [b] [50] y [c] [111]. Para asignar las posiciones se consideró el largo del segmento de ADN presente en el complejo proteico con relación a la ubicación del motivo de unión específico para cada proteína. Por ejemplo, la estructura de TBP (1C9B) contiene 7 pares de base previo al elemento TATA, por lo tanto la posición que se reporta esta corrida 7 bases corriente arriba. En todos los casos se reportan las posiciones corregidas. La distribución de los nucleosomas en regiones de cromatina se detalla en el texto.

Posiciones en la secuencia	Descripción	Estructura PDB
-3320 a -590	Contexto de cromatina en la célula blanco. Extremo 5'LTR	1KX5
[a] -408; 8662	Nuc 0	
[a] -4; 9082	Nuc 1	
[a] 266; 9352	Nuc 2	
470 a 8480	Genoma viral cromatinizado	
9556 a 1204	Contexto de cromatina en la célula blanco. Extremo 3'LTR	
[b] -254; 8832	NFAT	2093
[b] -173; 8913	USF	1AN4
[b] -142; 8944	LEF-1	2LEF
[b] -104; 8982	p50/RelA	3GUT
[b] -78; -68; -58; 9008; 9018; 9028	SP1	1TF3
[c] -33; 9053	RNAP II, complejo de iniciación	1C9B, 3K7A

El porcentaje de estructura conocida para proteínas humanas que participan en regulación de la transcripción del VIH-1 se obtuvo a partir de la base de datos *Uniprot* [112] y el *Protein Model Portal* [113]. La Tabla 2.2.2 lista las estructuras reportadas para las proteínas evaluadas.

TABLA 2.2.2 | Datos estructurales de proteínas relacionadas a la regulación de la transcripción en VIH-1. La información fue obtenida a partir de la base de datos *Uniprot* y el *Protein Model Portal*. Las estructuras que corresponden a la proteína evaluada poseen un 100% de identidad de secuencias. También se listan estructuras de proteínas identificadas como posibles moldes para el modelado estructural de algunas regiones en las proteínas de interés. [a] Proteína que une ADN.

Proteína	Código Uniprot	Tamaño	Estructuras reportadas	Región	% Identidad de secuencia
[a] p50	P19838	433	3GUT	41-352	100
			1SVC	2-365	100
			2O61	40-477	32
[a] p65	Q04206	551	3GUT	20-291	100
			1NFI	20-320	100
			3QXY	302-316	100
			2IW3	300-548	9
[a] NFAT1	Q13469	925	2O93	396-678	100
			2O61	410-886	19
[a] TBP	P20226	339	1C9B	159-337	100
			1NVP	159-339	100
			2DS2	55-118	22
[a] TFIIA(1)	P52655	376	1NVP	2-58, 303-376	100
			4DPV	48-290	7
[a] TFIIA(2)	P52657	109	1NVP	2-109	100
[a] TFIIB	Q00403	316	1C9B	110-316	100
			1RO4	1-60	100
[a] EST1	P14921	441	2NNY	280-441	100
			2KMD	29-138	98
			2QB0	64-271	24
[a] LEF1	Q9UJU2	399	2LEF	298-382	100
			3OUX	1-65	100
[a] SP1	P08047	785	1SP1	684-712	100
			1SP2	654-684	100
			1VA1	619-654	100
			1TF3	624-708	45
			1WXR	20-590	8
[a] USF1	P22415	310	1AN4	197-260	100
			1NKP	200-288	36
			1KZQ	5-196	14
[a] USF2	Q15853	346	2IW3	23-254	14
			1AN4	233-296	80
			1T3J	288-331	23
[a] AP1	P05412	331	1FOS	254-315	100
[a] LSF	Q12800	502	1WWV	308-422	29

TABLA 2.2.2 | Continuación.

Proteína	Código Uniprot	Tamaño	Estructuras reportadas	Región	% Identidad de secuencia
YY1	P25490	414	1UBD	293-414	100
p300	Q09472	2414	1L3E	323-423	100
			3I3J	1040-1161	100
			3BIY	1287-1666	100
			3IO2	1723-1836	100
			1Q0V	465-554	15
			1KDX	566-646	90
			1OQY	600-941	11
			1TOT	1663-1713	90
			2G8G	1854-2380	14
			1ZOQ	2050-2092	81
			1KBH	2045-2106	73
CTIP2	Q9C0K0	894	2COT	413-483	41
			2EBT	780-875	43
			2EM9	212-244	33
			1MEY	590-732	14
			1TF6	4-166	7
			1UBD	249-393	7
CBP	Q92793	2441	1WO6	133-138	100
			1LIQ	376-402	100
			2KWF	587-673	100
			3SVH	1081-1197	100
			2KJE	1763-1854	100
			1ZOQ	2065-2111	100
			1RDT	58-80	100
			1R8U	341-440	100
			3BIY	1323-1701	75
			3I3J	1074-1197	92
			1TOT	1699-1750	98
			2L14	2058-2116	98
			1HSS	2124-2239	12
			2G8G	657-1010	11
			2B5I	1892-2048	9
SUV39H1	O43463	412	3MTS	44-106	100
			2R3A	115-411	68
HP1	P83916	185	3F2U	20-73	100
			1GUW	110-185	100

TABLA 2.2.2 | Continuación.

Proteína	Código Uniprot	Tamaño	Estructuras reportadas	Región	% Identidad de secuencia
HDAC1	Q13547	482	1TYI	1-482	100
HDAC2	Q92769	488	3MAX	9-374	79
CDK9	P50750	372	3MI9	1-345	93
CyclinT1	O60563	726	2PK2	1-281	100
			1SJ8	370-657	9
pCAF	Q92831	832	1CM0	493-658	100
			1JM4	719-832	100
			3GG3	715-831	100

Resultados

El modelo estructural del genoma proviral de VIH-1 en estado de latencia se presenta en la Figura 2.2.2 A. En el se pueden apreciar dos zonas bien diferenciadas. Por un lado regiones de cromatina que abarcan gran parte del genoma viral y el contexto génico de la célula infectada, y por otro lado regiones mayormente accesibles a proteínas no histonicas en los extremos 5' y 3'LTR del virus. La accesibilidad de los LTR esta dada por el espaciamiento entre los nucleosomas Nuc0-Nuc1 (252 pares de base), y Nuc1-Nuc2 (110 pares de base), según fue observado por Verdin *et al.* [102]. Dicho espaciamiento esta generado por la unión de factores de transcripción a elementos de regulación, NF-κb y SP en el caso del VIH (ver Figura 1.0.1) [114], siendo una característica general a otros genomas [115,116]. La presencia de la RNAP II es un hecho que también condiciona la estructuración de los nucleosomas en torno a regiones de iniciación de la transcripción (Nuc1) [110,116]. En el caso del VIH-1, estudios de CHIP confirman su localización próxima al Nuc1 en estado activado e inactivado del virus [111]. La distancia de extremo a extremo de los LTR es de 170 nm, esta longitud solo sirve como referencia del tamaño del sistema ya que esta supeditada a la flexibilidad tanto en las regiones de cromatina como en las zonas más libres de proteínas. Ninguno de estos factores es tomado en cuenta de forma exhaustiva durante el presente modelado estructural. Es importante recordar que los segmentos de ADN que conectan entre complejos proteicos son modelados como hebras rectas en forma B canónica. Tampoco se realizó ningún proceso de refinado o minimización sistemático del modelo, solo se descartaron conformaciones que

por inspección visual presentaban solapamientos importantes entre moléculas. Por lo tanto los fenómenos estructurales que se observan son principalmente el resultado de deformaciones en regiones de unión a proteínas.

El modelo contiene más de 650 proteínas totales, de las cuales la mayoría corresponden a histonas que forman los nucleosomas y cuyo arreglo da lugar a la cromatina. Siendo esta última el componente estructural más importante del modelo cabe preguntarse si la conformación que surge del modelado es razonable. Actualmente existen dos modelos contrapuestos para explicar el arreglo espacial de los nucleosomas en las fibras de cromatina, los cuales se denominan modelo solenoide y modelo zigzag [117]. En el modelo solenoide la cadena de nucleosomas esta enrollada a modo de hélice alrededor de una cavidad interior de seis a ocho nucleosomas por vuelta y ~ 11 nm de paso. Los nucleosomas contiguos en secuencia interaccionan mediante apilamiento (Figura 2.2.3 A). En el modelo zigzag el ángulo de entrada y salida del ADN hace que dos nucleosomas consecutivos se orienten de forma parcialmente enfrentada. La posición de un tercer nucleosoma estará más cerca del primer nucleosoma que del segundo. La repetición de este patrón genera un entrecruzamiento que asemeja un zigzag, mientras que la estructura global de la cromatina tendrá un parecido a dos hélices enrolladas entre si (Figura 2.2.3 B). En este contexto, las interacciones de apilamiento se darán entre nucleosomas no contiguos en secuencia. Dado que ambos modelos cuentan con soporte experimental, no es claro cual representación prevalece en la célula [118]. Además las condiciones del entorno, la presencia de otras proteínas como la Histona 1 pueden influir en la conformación adoptada [119].

El método de ensamblado macromolecular empleado en este trabajo reproduce el modelo zigzag (Figura 2.2.2 B). Este arreglo surge como una propiedad intrínseca a la estructura del nucleosoma y los segmentos de ADN que los conectan. Las característica helicoidales del ADN hacen que distintas longitudes modifiquen la rotación entre nucleosomas, lo que cambia el arreglo espacial de la cromatina dando lugar a distintos grados de compactación (Figura 2.2.4). Estos efectos están cuantizados pues dependen de la torsión del ADN [120], existiendo restricciones topológicas en la combinación de ciertas longitudes del segmento de ADN que conecta los nucleosomas [121].

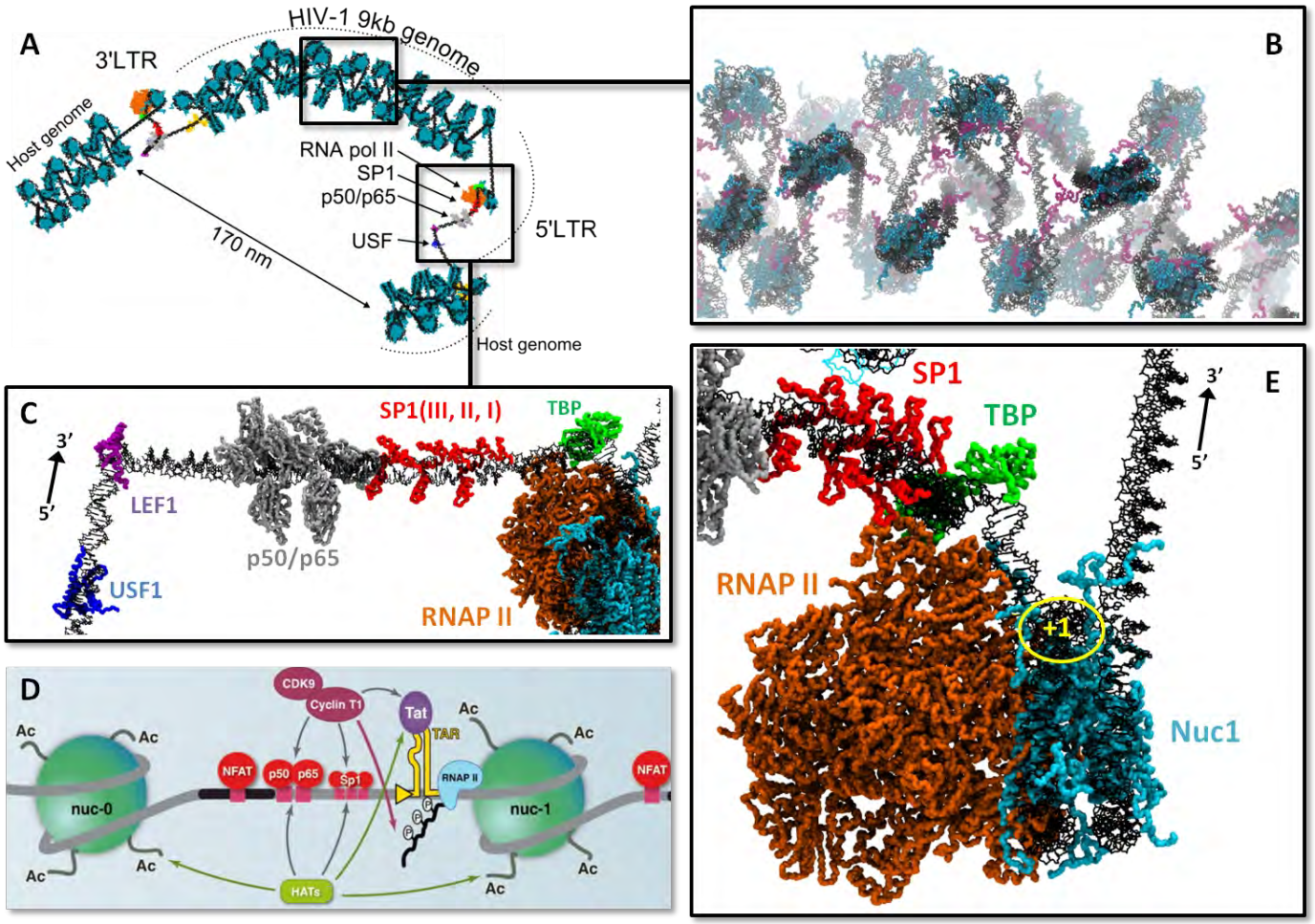

FIGURA 2.2.2 | Modelo macromolecular del provirus de VIH-1 en estado de latencia. A) Representación estructural resultante del ensamblado molecular. El filamento negro corresponde al ADN, en azul se pueden apreciar los nucleosomas, mientras que otras proteínas se señalan en la figura. Las zonas con gran empaquetamiento de los nucleosomas corresponden a regiones de cromatina. B) Ampliación de la región de cromatina enmarcada en la parte A de la figura. En azul se muestran los nucleosomas, mientras que en purpura se representa la histona 3.C) Ampliación del 5'LTR enmarcada en la parte A de la figura. Se indica la direccionalidad de la hebra de ADN. D) Figura adaptada de Trono et al. [68] en la cual se presenta esquemáticamente el modelo propuesto para la activación del VIH-1. El sitio de inicio de la transcripción se representa con un triángulo amarillo. E) Detalles de la región cercana a la RNAP II, en la que además se pueden apreciar las proteínas SP1, TBP y el nucleosoma 1 (Nuc1). El circulo amarillo muestra la localización del sitio de iniciación de la transcripción (+1). Se indica la direccionalidad de la hebra de ADN.

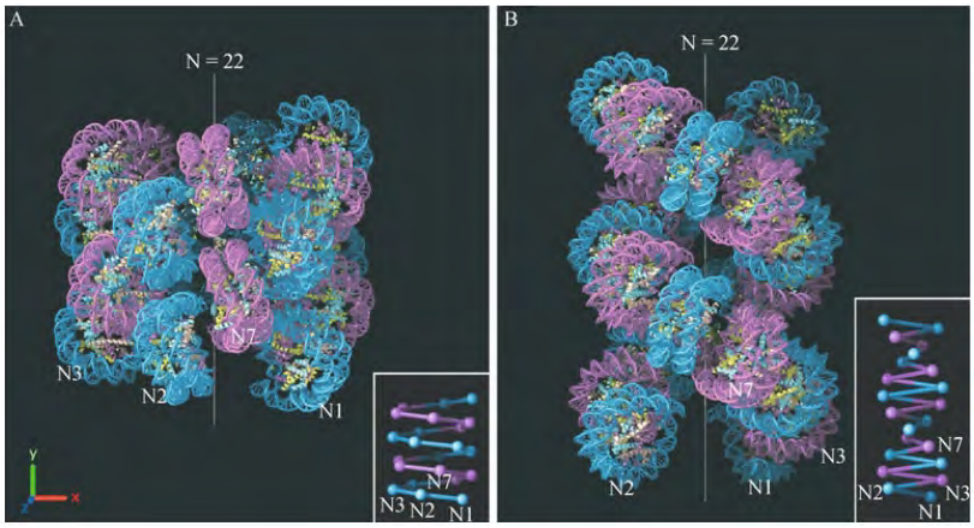

FIGURA 2.2.3 | Modelo solenoide (A) y zigzag (B) para una fibra de cromatina de 30 nm compuesta por 22 nucleosomas (N=22). Se señalan las posiciones del primer, segundo, tercer y séptimo nucleosoma (N1, N2, N3 y N7). Los esquemas interiores clarifican el arreglo entre los nucleosomas para cada modelo. Figura adaptada de [117].

Según observaron Woodcock *et al.* pequeños cambios en la longitud del ADN pueden generar grandes cambios de curvatura e incluso quiebres en el filamento de cromatina [122]. Esta característica también puede ser modelada en el presente ensamblado. El aumento o disminución de 1 par de bases en 6 segmentos de ADN es capaz de generar un cambio notable en toda la fibra (Figura 2.2.2 A). Puesto que el tamaño del ADN inter-nucleosoma es menor que la longitud de persistencia para el ADN (50 nm, 150 pares de base [123]) la flexibilidad de la doble hebra tendrá menos impacto sobre la conformación de la cromatina que el cambio en la longitud del segmento [122]. Por lo tanto aproximar el ADN inter-nucleosoma con un segmento recto es razonable mientras este tenga una extensión menor a la longitud de persistencia.

La conformación de la cromatina genera restricciones espaciales en la interacción del ADN o los nucleosomas con otras proteínas. Como se puede observar de la Figura 2.2.2 B los segmentos N terminales de la Histona 3 (H3) se encuentran orientados principalmente hacia adentro del entramado de nucleosomas. Este ordenamiento también se ha observado en el modelo solenoide [124]. Según se ha visto en la sección anterior del presente trabajo, la proteína HP1 juega un rol importante en la expresión del genoma de VIH. En particular HP1 interacciona con H3, por lo que para hacerlo deberá moverse a través de los intersticios de la fibra de cromatina. Recientemente se ha

propuesto la oligomerización de HP1 como un mecanismo para interconectar nucleosomas, favoreciendo la compactación y estabilidad de la cromatina [89]. En ese modelo, el dominio Chromo de HP1 es capaz no solo de unirse a H3 sino que puede dimerizar con otro dominio Chromo de una HP1 ligada a un nucleosoma cercano, aumentando el vínculo entre ambos nucleosomas. Así mismo, los autores observaron una fuerte relación entre la longitud del ADN que espacia los nucleosomas y la capacidad de oligomerización de HP1 [89]. Según nuestro modelo estructural esto puede explicarse por el espacio excluido disponible para acceder a interacciones específicas, el cual dependerá de la estructura de la cromatina (ver Figura 2.2.4). De esta forma la organización de la red de HP1 estará influenciada a su vez por el largo del segmento de ADN inter-nucleosoma.

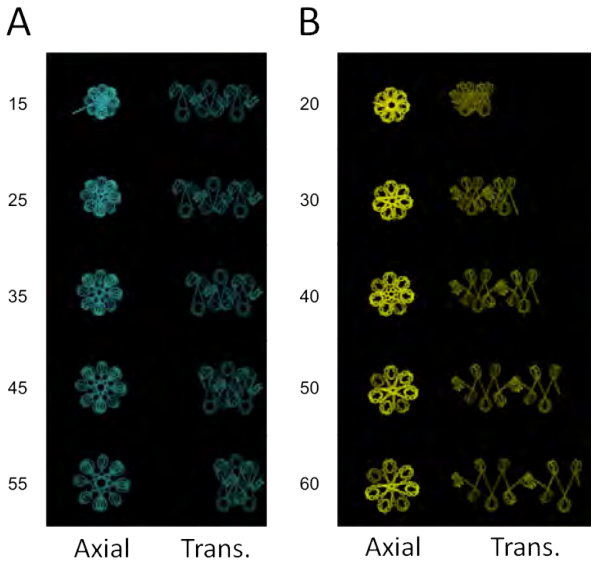

FIGURA 2.2.4 | Estructura de la cromatina en función de la longitud del segmento de ADN inter-nocleosoma. Modelos zigzag obtenidos para el ensamblado de 12 nucleosomas sobre un segmento de ADN. A) Vistas axial y transversal de la fibra de cromatina, se omiten las histonas por claridad. A la izquierda se muestra la longitud en pares de base del ADN que conecta los nucleosomas en cada modelo. Estas longitudes siguen la regla 10*n*+5, donde *n* es un número entero mayor a cero. B) Ídem que la parte A de la figura, pero en este caso las longitudes del ADN que conecta los nucleosomas siguen la regla 10*n*, con *n* mayor a 1.

La región comprendida entre los nucleosomas Nuc0 y Nuc1 (posiciones de -300 a +10) es de particular interés para la regulación del VIH-1. Como muestran ensayos de protección varias proteínas están unidas de forma constitutiva a elementos de regulación próximos al inicio de la transcripción [125]. En particular los sitios de unión SP1 y NF-κb se encuentran siempre

ocupados independientemente del estado de activación del virus. En el sitio NF-κb se pueden ligar factores como NFAT1 [126], p50 o p65, dando lugar a distintos estados de expresión [49,127]. La proximidad que muestra el modelo estructural entre proteínas humanas como SP1 (sitio III) y p50/p65 Figura 2.2.2 C, es una consecuencia específica de la secuencia del ADN viral. Esto plantea posibles superficies de contacto que podrían ser explotadas en el diseño de drogas. Resulta interesante contrastar el modelo estructural generado en esta Tesis con esquemas provenientes de la biología molecular. En el esquema empleado como ejemplo se muestra como la RNAP II en estado de elongación se ubica entre el sitio de iniciación de la transcripción y el nucleosoma Nuc1 (Figura 2.2.2 D). Los primeros 50 a 80 nucleótidos transcriptos por la RNAP II forman el elemento TAR, este corresponde a un segmento de ARN capaz de adoptar una estructura en forma de horquilla que reconoce y liga al factor viral Tat. La proteína Tat actúa a su vez reclutando factores de transcripción humanos como el complejo formado por la quinasa CDK9 y la ciclina T1, que potencia la transcripción. Este esquema presenta algunas discrepancias respecto al modelo estructural presentado en esta Tesis. De acuerdo a nuestro modelo, el nucleosomas Nuc1 cubre la secuencia desde -2 a +145, ubicándose muy próximo al sitio de unión corriente abajo para NFAT (bases +162 a +170 [128]). Una vista detallada de la región RNAP II muestra que el sitio (+1) se encuentra muy cerca del nucleosoma Nuc1 (Figura 2.2.2 E). Para que la RNAP II tenga espacio suficiente para transcribir el elemento TAR es necesario mover al menos 80 pares de bases corriente abajo el nucleosoma Nuc1, con lo cual la secuencia de unión a NFAT quedaría oculta dentro de este último. A su vez esta nueva ubicación haría que el nucleosoma Nuc1 se posicione muy próximo (a 30 pares de bases) del nucleosoma Nuc2. Por lo tanto el esquema tal cual se presenta en la Figura 2.2.2 D no sería físicamente plausible. Como se aprecia en la Figura 2.2.2 E, el nucleosoma Nuc1 constituye una barrera física para la RNAP II, para que la transcripción sea posible es necesario levantar esta barrera [9,129]. Datos experimentales muestran que el nucleosoma sufre modificaciones epigenéticas durante la activación del virus a su vez, complejos de remodelación de la cromatina también participan en el proceso que da lugar a la transcripción [130,131]. Estos son ejemplos de como el modelo estructural puede ayudar a comprender e interpretar los datos experimentales.

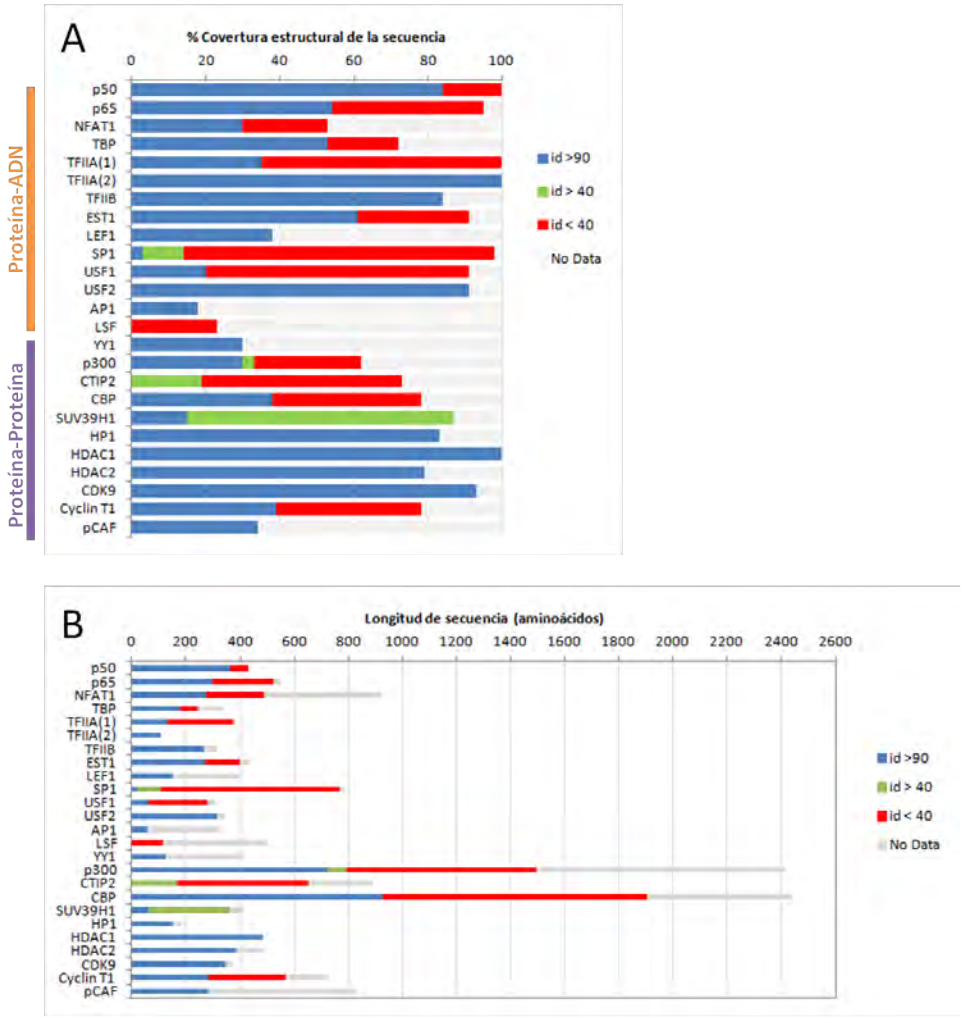

FIGURA 2.2.5 | Datos estructurales disponibles para distintas proteínas humanas implicadas en la regulación del VIH-1. A) Porcentaje de la secuencia para la cual existen estructuras resueltas en la base de datos PDB con identidad mayor a 90 (id>90), entre 40 y 90 (id>40), menor de 40 (id<40) o sin datos (No Data). Los resultados para cada proteína fueron obtenidos a partir de la Tabla 2.2.2. No se consideran posibles regiones desordenadas en las proteínas o bucles que conecten dominios. En el margen izquierdo del gráfico se indican cuales son las proteínas que median interacciones principalmente del tipo proteína-ADN y proteína-proteína. B) ídem a la parte A de la figura, pero los valores se reportan en términos de longitud de secuencia cubierta.

La estrategia empleada para ensamblar el modelo del genoma proviral de VIH-1 presenta varias limitaciones. Como se menciono antes las principales deformaciones del ADN surgen de la interacción con proteínas. Aun con esta carencia es notable como se puede modelar de forma razonable la conformación de la cromatina. Lo cual indica que en algún punto los fragmentos pueden contener información de como se relacionan con otros. Otra limitación en el modelo es emplear principalmente proteínas que se unen de forma directa al ADN. Es importante tener en cuenta que existe un conjunto de interacciones proteína-proteína acoplado a las proteínas de unión al ADN que cumple un rol muy importante en la regulación (Ver Tabla 2.2.2). Sin

embargo el modelado de dichas interacciones en el contexto del presente ensamblado macromolecular es mucho más complejo y requiere de la combinación de otras estrategias computacionales así como mayor información estructural referente a las proteínas que participan en las interacciones.

Independientemente de la estrategia de ensamblado empleada, un factor importante que limita el modelado es la información estructural disponible para cada proteína. Como muestra la Figura 2.2.5 A existen proteínas para las cuales el nivel de conocimiento estructural supera el 80% de la secuencia, mientras que para otras apenas si llega al 20% o menos. Uno de estos caso es la proteína SP1, de la cual solo se conoce el dominio de unión al ADN. Para el resto de su secuencia (aprox. 800 residuos) el porcentaje de identidad con estructuras conocidas es menor a 40% Figura 2.2.5 B. Hacer modelos por homología en base a esta baja identidad de secuencia requiere un cuidado particular. Al no considerar el espacio que ocupan los segmentos no modelados de las proteínas se esta subestimando el volumen excluido que existe para interaccionar. Esto también, dificulta el modelado de interacciones con otras proteínas. Por ejemplo, en el caso de SP1 implica desconocer los dominios implicados en la oligomerizarción y formación de bucles en el ADN [132]. Otro caso lo constituyen las proteínas p300 y CBP para las cuales el 40% de la secuencia cuenta con datos estructurales, sin embargo el porcentaje restante del cual no hay datos, equivale al doble del tamaño de SP1 Figura 2.2.5. Diferente es lo que ocurre con proteínas como HP1, SUV39H1, HDAC1, HDAC2 o CDK9, para las cuales se conoce gran parte de su estructura, pero poco del complejo multiproteico que integran. Por último, las proteínas p50 y p65 constituyen un ejemplo muy interesante pues se conoce entre el 60-80% (aprox. 400 residuos) de su estructura mientras que se desconoce la conformación de al máximo 100 residuos Figura 2.2.5. Por lo tanto el volumen que ambas proteínas representan en el modelo estructural (Figura 2.2.2 C) no será muy distinto del de la proteína completa. El complejo p50/p65 debe constituir una plataforma de anclaje para otras proteínas que permitan mediar interacciones a larga distancia con la RNAP II, de otra manera no serían posibles estas interacciones.

Conclusiones

Se presentó por primera vez un modelo estructural del genoma proviral de VIH-1 en estado de latencia. Si bien la aproximación empleada para el ensamblado macromolecular es rudimentario e implica la concatenación de estructuras de complejos ADN-proteína sobre una fibra de ADN, ésta estrategia resulto ser razonable para generar un modelo estructural inicial. A su vez permitió combinar información estructural y bioquímica procedente de diferentes enfoques. En este sentido el modelo resume de alguna manera todos los conocimientos adquiridos sobre varios complejos proteicos, acumulados durante más de 20 años y donde cada componente constituye un objeto de estudio en si mismo enmarcado en la regulación génica eucariota.

Las interacciones proteína-proteína y la flexibilidad de los segmentos que conectan los complejos proteicos constituyen un desafío a incorporar en el modelo. A pesar de esto, la estructura de la cromatina fue reproducida razonablemente. También se pudieron realizar algunas observaciones sobre la interacción de HP1 con H3, SP1 con p50/p65 y RNAP II con el nucleosoma Nuc1.

Si bien la cantidad de dominios de plegamiento estructural parece estar convergiendo y cada vez se descubren menos tipos nuevos [133], existe desconocimiento estructural en la secuencia de varias proteínas. Este es un aspecto muy importante a considerar a la hora de realizar cualquier modelado estructural.

Todos estos factores hacen que la resolución final del modelo sea difícil de estimar, si bien el nivel de descripción es pseudoatómico (en el sentido que están representados los átomos), la precisión de la estructura dependerá de la región observada. A pesar de ello, el modelo permitió acceder a un conocimiento nuevo sobre al tamaño relativo, orientación y ubicación de los componentes, así como resulto útil para identificar zonas de posible contacto entre proteínas.

CONSIDERACIONES FINALES

En el presente trabajo de Tesis se abordaron dos temas de gran interés para el VIH-1 como ser los mecanismos de transcripción y represión del genoma proviral. Para la ejecución del mismo se emplearon técnicas de modelado *in silico,* cubriendo un amplio espectro de escalas espaciales (desde proteínas al genoma) y temporales (de nanosegundos a microsegundos). Como parte del trabajo se desarrollo, validó y aplicó un modelo simplificado de ADN a un problema de interés biológico. Este modelo abre un campo completamente nuevo e innovador de posibilidades desde el punto de vista computacional.

Con respecto a los mecanismos de transcripción se exploró el rol que puede cumplir el ADN como medio para la transducción de señales. Se observó que diferentes secuencias son capaces de reaccionar de modo distinto a la unión de factores de transcripción cambiando sus patrones estructurales a gran distancia del sitio de unión. Este efecto es apreciable inclusive en secuencias que difieren en muy pocos nucleótidos. Tales resultados plantean un ajuste fino tanto de los nucleótidos que se unen a proteínas como los que no. Dentro del contexto de promotores virales, como el del VIH-1, se abre la interrogante sobre el impacto que pueden tener diferentes polimorfismos en la transcripción.

En relación a los mecanismos de represión del VIH-1, se investigó la interacción de dos isoformas de HP1 con la histona 3. De este estudio se obtuvieron determinantes estructurales característicos a cada isoforma. Es de destacar que los residuos comprometidos en comportamientos isoforma específica se ubican en regiones poco estructuradas al N terminal del dominio Chromo de HP1, lo que concuerda con algunos datos experimentales recientes. Por último se proporcionó una visión estructural de la organización del genoma proviral en el estado de latencia. Con esta aproximación se intentó tender un puente entre la escala atomístico de los fenómenos y la observación macroscópica de los experimentos.

REFERENCIAS

[1] Rambaut, A., Posada, D., Crandall, K.A., and Holmes, E.C. (2004) The causes and consequences of HIV evolution. *Nat Rev Genet*. 5: 52-61.

[2] Gilbert, P.B., McKeague, I.W., Eisen, G., Mullins, C., Guéye-NDiaye, A., Mboup, S., and Kanki, P.J. (2003) Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Stat Med*. 22: 573-593.

[3] Liu, J., Bartesaghi, A., Borgnia, M.J., Sapiro, G., and Subramaniam, S. (2008) Molecular architecture of native HIV-1 gp120 trimers. *Nature*. 455: 109-113.

[4] Hare, S., Gupta, S.S., Valkov, E., Engelman, A., and Cherepanov, P. (2010) Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature*. 464: 232-236.

[5] Ott, M., Geyer, M., and Zhou, Q. (2011) The control of HIV transcription: keeping RNA polymerase II on track. *Cell Host Microbe*. 10: 426-435.

[6] Trono, D., Van Lint, C., Rouzioux, C., Verdin, E., Barré-Sinoussi, F., Chun, T.-W., and Chomont, N. (2010) HIV persistence and the prospect of long-term drug-free remissions for HIV-infected individuals. *Science*. 329: 174-180.

[7] Pearson, R., Kim, Y.K., Hokello, J., Lassen, K., Friedman, J., Tyagi, M., and Karn, J. (2008) Epigenetic silencing of human immunodeficiency virus (HIV) transcription by formation of restrictive chromatin structures at the viral long terminal repeat drives the progressive entry of HIV into latency. *J Virol*. 82: 12291-12303.

[8] Hakre, S., Chavez, L., Shirakawa, K., and Verdin, E. (2011) Epigenetic regulation of HIV latency. *Curr Opin HIV AIDS*. 6: 19-24.

[9] Marcello, A. (2006) Latency: the hidden HIV-1 challenge. *Retrovirology*. 3: 7.

[10] Juven-Gershon, T., and Kadonaga, J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol*. 339: 225-229.

[11] Kilareski, E.M., Shah, S., Nonnemacher, M.R., and Wigdahl, B. (2009) Regulation of HIV-1 transcription in cells of the monocyte-macrophage lineage. *Retrovirology*. 6: 118.

[12] Monod, J., Wyman, J., and Changeux, J.P. (1965) ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *J Mol Biol*. 12: 88-118.

[13] Popovych, N., Sun, S., Ebright, R.H., and Kalodimos, C.G. (2006) Dynamically driven protein allostery. *Nat Struct Mol Biol*. 13: 831-838.

[14] Chaires, J.B. (2008) Allostery: DNA does it, too. *ACS Chem Biol*. 3: 207-209.

[15] de Arellano, E.R., Alcamí, J., López, M., Soriano, V., and Holguín, A. (2010) Drastic decrease of transcription activity due to hypermutated long terminal repeat (LTR) region in different HIV-1 subtypes and recombinants. *Antiviral Res*. 88: 152-159.

[16] Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. 79: 233-269.

[17] Sadowski, I., Lourenco, P., and Malcolm, T. (2008) Factors controlling chromatin organization and nucleosome positioning for establishment and maintenance of HIV latency. *Curr HIV Res*. 6: 286-295.

[18] du Chéné, I., Basyuk, E., Lin, Y.-L., Triboulet, R., Knezevich, A., Chable-Bessia, C., Mettling, C., Baillat, V., Reynes, J., Corbeau, P., Bertrand, E., Marcello, A., Emiliani, S., Kiernan, R., and Benkirane, M. (2007) Suv39H1 and HP1gamma are responsible for chromatin-mediated HIV-1 transcriptional silencing and post-integration latency. *EMBO J*. 26: 424-435.

[19] Mateescu, B., Bourachot, B., Rachez, C., Ogryzko, V., and Muchardt, C. (2008) Regulation of an inducible promoter by an HP1beta-HP1gamma switch. *EMBO Rep*. 9: 267-272.

[20] Abbas, W., and Herbein, G. (2012) Molecular Understanding of HIV-1 Latency. *Adv Virol*. 2012: 574967.

[21] Leach, A.A.Leach, A.A. (Ed.) (2001) *Molecular modelling: principles and applications* Addison-Wesley Longman Ltd.

[22] Swope, W.C., Andersen, H.C., Berens, P.H., and Wilson, K.R. (1982) A computer-simulation method for the calculation of equilibrium-constants for the formation of physical clusters of molecules: Application to small water clusters *J Chem Phys*. 76: 637-649.

[23] AMBER (2012) Assisted Model Building with Energy Refinement.. University of California, San Francisco. Available http://ambermd.org/.

[24] CHARMM (2012) Chemistry at HARvard Macromolecular Mechanics.. Available http://www.charmm.org/.

[25] GROMOS (2012) GROningen MOlecular Simulation.. Available http://www.gromos.net/.

[26] Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996) Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids *J Am Chem Soc*. 118: 11225-11236.

[27] Zwier, M.C., and Chong, L.T. (2010) Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol*. 10: 745-752.

[28] Vendruscolo, M., and Dobson, C.M. (2011) Protein dynamics: Moore's law in molecular biology. *Curr Biol*. 21: R68-R70.

[29] Nosé, S. (1984) A molecular dynamics method for simulations in the canonical ensemble *Mol Phys*. 52: 255-268.

[30] Hoover, W.G. (1985) Canonical dynamics: equilibrium phase-space distributions *Phys Rev A*. 31: 1695-1697.

[31] Berendsen, H.J.C., Postma, J.P.M., DiNola, A., and Haak, J.R. (1984) Molecular dynamics with coupling to an external bath *J Chem Phys*. 81: 3684-3690.

[32] Parrinello, M., and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys*. 52: 7182-7190.

[33] Nosé, S., and Klein, M.L. (1983) Constant pressure molecular dynamics for molecular systems *Mol Phys*. 50: 1055-1076.

[34] Jorgensen, W.L., Chandrasekaran, R., Madura, J.D., Impey, R.W., and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water *J Chem Phys*. 79: 926-935.

[35] Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., and Hermans, J.Pullman, B. (Ed.) (1981) *Interaction models for water in relation to protein hydration. In: Intermolecular Forces.* Dordrecht: Reidel.

[36] Feig, M., and Brooks, 3rd, C.L. (2004) Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr Opin Struct Biol*. 14: 217-224.

[37] Roux, B., and Simonson, T. (1999) Implicit solvent models. *Biophys Chem*. 78: 1-20.

[38] Hawkins, G.D., J., C.C., and Truhlar, D.G. (1995) Pairwise solute descreening of solute charges from a dielectric medium. *Chem Phys Lett*. 246: 122-129.

[39] Hawkins, D., Cramer, C., and Truhlar, D. (1996) Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J Phys Chem A*. 100: 19824-19839.

[40] Qui, D., Shenkin, P., Hollinger, F., and Still, W. (1997) The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J Phys Chem A*. 101: 3005-3014.

[41] Tom, D., Darrin, Y., and Lee, P. (1993) Particle mesh Ewald: An N log(N) method for Ewald sums in large systems *J Chem Phys*. 98: 10089-10092.

[42] Voth, G.A.Voth, G.A. (Ed.) (2009) *Coarse-graining of condensed phase and biomolecular systems* CRC Press / Taylor & Francis Group, New York.

[43] Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., and de Vries, A.H. (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B*. 111: 7812-7824.

[44] Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., and Marrink, S.J. (2008) The MARTINI Coarse-Grained Force Field: Extension to Proteins *J Chem Theory Comput*. 4: 819-834.

[45] Dans, P.D., Zeida, A., Machado, M.R., and Pantano, S. (2010) A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics *J Chem Theory Comput*. 6: 1711-1725.

[46] Darré, L., Machado, M.R., Dans, P.D., Herrera, F.E., and Pantano, S. (2010) Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *J Chem Theory Comput*. 6: 3793-3807.

[47] Zeida, A., Machado, M.R., Dans, P.D., and Pantano, S. (2012) Breathing, bubbling and bending: DNA flexibility from multi microseconds simulations. *Phys Rev E*. [Accepted].

[48] Dans, P.D., Darré, L., Machado, M.R., Zeida, A., and Pantano, S.Villà-Freixa, J. (Ed.) (en prensa) *Coarse grain potential: a model for DNA in implicit and explicit solvent. En "A course on biomolecular simulations"* Huygens.

[49] Colin, L., and Van Lint, C. (2009) Molecular control of HIV-1 postintegration latency: implications for the development of new therapeutic strategies. *Retrovirology*. 6: 111.

[50] Pereira, L.A., Bentley, K., Peeters, A., Churchill, M.J., and Deacon, N.J. (2000) A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res*. 28: 663-668.

[51] Gertz, J., Siggia, E.D., and Cohen, B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*. 457: 215-218.

[52] Mogno, I., Vallania, F., Mitra, R.D., and Cohen, B.A. (2010) TATA is a modular component of synthetic promoters. *Genome Res*. 20: 1391-1397.

[53] Schurr, J.M., Delrow, J.J., Fujimoto, B.S., and Benight, A.S. (1997) The question of long-range allosteric transitions in DNA. *Biopolymers*. 44: 283-308.

[54] Travers, A.A. (2004) The structural basis of DNA flexibility. *Philos Transact A Math Phys Eng Sci*. 362: 1423-1438.

[55] Altan-Bonnet, G., Libchaber, A., and O., K. (2003) Bubble Dynamics in Double-Stranded DNA *Phys Rev Lett*. 90: 138101.

[56] Nikolov, D.B., and Burley, S.K. (1997) RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A*. 94: 15-22.

[57] Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*. 461: 186-192.

[58] Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature*. 365: 512-520.

[59] Nikolov, D.B., Chen, H., Halay, E.D., Hoffman, A., Roeder, R.G., and Burley, S.K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc Natl Acad Sci U S A*. 93: 4862-4867.

[60] Tsai, F.T., and Sigler, P.B. (2000) Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J*. 19: 25-36.

[61] Bleichenbacher, M., Tan, S., and Richmond, T.J. (2003) Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J Mol Biol*. 332: 783-793.

[62] Savelyev, A. (2012) Do monovalent mobile ions affect DNA's flexibility at high salt content? *Phys Chem Chem Phys*. 14: 2250-2254.

[63] Knotts, 4th, T.A., Rathore, N., Schwartz, D.C., and de Pablo, J.J. (2007) A coarse grain model for DNA. *J Chem Phys*. 126: 084901.

[64] Savin, A.V., Mazo, M.A., Kikot, I.P., Manevitch, L.I., and Onufriev, A.V. (2011) Heat conductivity of the DNA double helix *Phys Rev B*. 83: 15pp.

[65] Nielsen, S.O., Bulo, R.E., Moore, P.B., and Ensing, B. (2010) Recent progress in adaptive multiscale molecular dynamics simulations of soft matter. *Phys Chem Chem Phys*. 12: 12401-12414.

[66] Machado, M.R., Dans, P.D., and Pantano, S. (2011) A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys Chem Chem Phys*. 13: 18134-18144.

[67] GROMACS (2012) Groningen Machine for Chemical Simulations.. Available http://www.gromacs.org/.

[68] Juven-Gershon, T., Cheng, S., and Kadonaga, J.T. (2006) Rational design of a super core promoter that enhances gene expression. *Nat Methods*. 3: 917-922.

[69] Alexandrov, B.S., Gelev, V., Yoo, S.W., Alexandrov, L.B., Fukuyo, Y., Bishop, A.R., Rasmussen, K.Ø., and Usheva, A. (2010) DNA dynamics play a role as a basal transcription factor in the positioning and

regulation of gene transcription initiation. *Nucleic Acids Res*. 38: 1790-1795.

[70] Davis, N.A., Majee, S.S., and Kahn, J.D. (1999) TATA box DNA deformation with and without the TATA box-binding protein. *J Mol Biol*. 291: 249-265.

[71] Wang, J., Cieplak, P., and Kollman, P.A. (2000) How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J Comput Chem*. 21: 1049-1074.

[72] Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, 3rd, T.E., Laughton, C.A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J*. 92: 3817-3829.

[73] Srinivasan, J., Trevathan, M.W., Beroza, P., and Case, D.A. (1999) Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects *Theor Chem Acc*. 101: 426-434.

[74] Alexandrov, B.S., Gelev, V., Yoo, S.W., Bishop, A.R., Rasmussen, K.Ø., and Usheva, A. (2009) Toward a detailed description of the thermally induced dynamics of the core promoter. *PLoS Comput Biol*. 5: e1000313.

[75] Faiger, H., Ivanchenko, M., Cohen, I., and Haran, T.E. (2006) TBP flanking sequences: asymmetry of binding, long-range effects and consensus sequences. *Nucleic Acids Res*. 34: 104-119.

[76] Fairley, J.A., Evans, R., Hawkes, N.A., and Roberts, S.G.E. (2002) Core promoter-dependent TFIIB conformation and a role for TFIIB conformation in transcription start site selection. *Mol Cell Biol*. 22: 6697-6705.

[77] Dikstein, R. (2011) The unexpected traits associated with core promoter elements. *Transcription*. 2: 201-206.

[78] Le Douce, V., Herbein, G., Rohr, O., and Schwartz, C. (2010) Molecular mechanisms of HIV-1 persistence in the monocyte-macrophage lineage. *Retrovirology*. 7: 32.

[79] Cui, P., Zhang, L., Lin, Q., Ding, F., Xin, C., Fang, X., Hu, S., and Yu, J. (2010) A novel mechanism of epigenetic regulation: nucleosome-space occupancy. *Biochem Biophys Res Commun*. 391: 884-889.

[80] Guillemette, B., Drogaris, P., Lin, H.-H.S., Armstrong, H., Hiragami-Hamada, K., Imhof, A., Bonneil, E., Thibault, P., Verreault, A., and Festenstein, R.J. (2011) H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation. *PLoS Genet*. 7: e1001354.

[81] Grewal, S.I.S., and Moazed, D. (2003) Heterochromatin and epigenetic control of gene expression. *Science*. 301: 798-802.

[82] Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., and Patel, D.J. (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol*. 14: 1025-1040.

[83] Burnett, J.C., Miller-Jensen, K., Shah, P.S., Arkin, A.P., and Schaffer, D.V. (2009) Control of stochastic gene expression by host factors at the HIV promoter. *PLoS Pathog*. 5: e1000260.

[84] Lomberk, G., Wallrath, L., and Urrutia, R. (2006) The Heterochromatin Protein 1 family. *Genome Biol*. 7: 228.

[85] Dialynas, G.K., Terjung, S., Brown, J.P., Aucott, R.L., Baron-Luhr, B., Singh, P.B., and Georgatos, S.D. (2007) Plasticity of HP1 proteins in mammalian cells. *J Cell Sci*. 120: 3415-3424.

[86] Minc, E., Courvalin, J.C., and Buendia, B. (2000) HP1gamma associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. *Cytogenet Cell Genet*. 90: 279-284.

[87] Nielsen, A.L., Oulad-Abdelghani, M., Ortiz, J.A., Remboutsika, E., Chambon, P., and Losson, R. (2001) Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. *Mol Cell*. 7: 729-739.

[88] Smothers, J.F., and Henikoff, S. (2000) The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr Biol*. 10: 27-30.

[89] Canzio, D., Chang, E.Y., Shankar, S., Kuchenbecker, K.M., Simon, M.D., Madhani, H.D., Narlikar, G.J., and Al-Sady, B. (2011) Chromodomain-mediated oligomerization of HP1 suggests a

nucleosome-bridging mechanism for heterochromatin assembly. *Mol Cell*. 41: 67-81.

[90] Hiragami, K., and Festenstein, R. (2005) Heterochromatin protein 1: a pervasive controlling influence. *Cell Mol Life Sci*. 62: 2711-2726.

[91] Zarebski, M., Wiernasz, E., and Dobrucki, J.W. (2009) Recruitment of heterochromatin protein 1 to DNA repair sites. *Cytometry A*. 75: 619-625.

[92] Luijsterburg, M.S., Dinant, C., Lans, H., Stap, J., Wiernasz, E., Lagerwerf, S., Warmerdam, D.O., Lindh, M., Brink, M.C., Dobrucki, J.W., Aten, J.A., Fousteri, M.I., Jansen, G., Dantuma, N.P., Vermeulen, W., Mullenders, L.H.F., Houtsmuller, A.B., Verschure, P.J., and van Driel, R. (2009) Heterochromatin protein 1 is recruited to various types of DNA damage. *J Cell Biol*. 185: 577-586.

[93] Shapiro, E., Huang, H., Ruoff, R., Lee, P., Tanese, N., and Logan, S.K. (2008) The heterochromatin protein 1 family is regulated in prostate development and cancer. *J Urol*. 179: 2435-2439.

[94] Poleshko, A., Palagin, I., Zhang, R., Boimel, P., Castagna, C., Adams, P.D., Skalka, A.M., and Katz, R.A. (2008) Identification of cellular proteins that maintain retroviral epigenetic silencing: evidence for an antiviral response. *J Virol*. 82: 2313-2323.

[95] Machado, M.R., Dans, P.D., and Pantano, S. (2010) Isoform-specific determinants in the HP1 binding to histone 3: insights from molecular simulations. *Amino Acids*. 38: 1571-1581.

[96] Nielsen, P.R., Nietlispach, D., Mott, H.R., Callaghan, J., Bannister, A., Kouzarides, T., Murzin, A.G., Murzina, N.V., and Laue, E.D. (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature*. 416: 103-107.

[97] Jacobs, S.A., and Khorasanizadeh, S. (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science*. 295: 2080-2083.

[98] Fischle, W., Wang, Y., Jacobs, S.A., Kim, Y., Allis, C.D., and Khorasanizadeh, S. (2003) Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev*. 17: 1870-1881.

[99] Hiragami-Hamada, K., Shinmyozu, K., Hamada, D., Tatsu, Y., Uegaki, K., Fujiwara, S., and Nakayama, J.-I. (2011) N-terminal phosphorylation of HP1alpha promotes its chromatin binding. *Mol Cell Biol*. 31: 1186-1200.

[100] Vakoc, C.R., Mandat, S.A., Olenchock, B.A., and Blobel, G.A. (2005) Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell*. 19: 381-391.

[101] Smale, S.T., and Fisher, A.G. (2002) Chromatin structure and gene regulation in the immune system. *Annu Rev Immunol*. 20: 427-462.

[102] Verdin, E., Paras, Jr, P., and Van Lint, C. (1993) Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *EMBO J*. 12: 3249-3259.

[103] Choudhary, S.K., and Margolis, D.M. (2011) Curing HIV: Pharmacologic approaches to target HIV-1 latency. *Annu Rev Pharmacol Toxicol*. 51: 397-418.

[104] Huang, S.-Y., and Zou, X. (2010) Advances and challenges in protein-ligand docking. *Int J Mol Sci*. 11: 3016-3034.

[105] Werner, T., Morris, M.B., Dastmalchi, S., and Church, W.B. (2012) Structural modelling and dynamics of proteins for insights into drug interactions. *Adv Drug Deliv Rev*. 64: 323-343.

[106] Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Sali, A., and Rout, M.P. (2007) The molecular architecture of the nuclear pore complex. *Nature*. 450: 695-701.

[107] Panne, D., Maniatis, T., and Harrison, S.C. (2007) An atomic model of the interferon-beta enhanceosome. *Cell*. 129: 1111-1123.

[108] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010) A three-dimensional model of the yeast genome. *Nature*. 465: 363-367.

[109] Bushnell, D.A., Westover, K.D., Davis, R.E., and Kornberg, R.D. (2004) Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. *Science*. 303: 983-988.

[110] Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 132: 887-898.

[111] Perkins, K.J., Lusic, M., Mitar, I., Giacca, M., and Proudfoot, N.J. (2008) Transcription-dependent gene looping of the HIV-1 provirus is dictated by recognition of pre-mRNA processing signals. *Mol Cell*. 29: 56-68.

[112] UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 40: D71-D75.

[113] Arnold, K., Kiefer, F., Kopp, J., Battey, J.N.D., Podvinec, M., Westbrook, J.D., Berman, H.M., Bordoli, L., and Schwede, T. (2009) The Protein Model Portal. *J Struct Funct Genomics*. 10: 1-8.

[114] Widłak, P., and Garrard, W.T. (1998) Nucleosomes and regulation of gene expression. Structure of the HIV-1 5'LTR. *Acta Biochim Pol*. 45: 209-219.

[115] Goh, W.S., Orlov, Y., Li, J., and Clarke, N.D. (2010) Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Comput Biol*. 6: e1000649.

[116] Jansen, A., and Verstrepen, K.J. (2011) Nucleosome positioning in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*. 75: 301-320.

[117] Chen, P., and Li, G. (2010) Dynamics of the higher-order structure of chromatin. *Protein Cell*. 1: 967-971.

[118] Szerlong, H.J., and Hansen, J.C. (2011) Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem Cell Biol*. 89: 24-34.

[119] Routh, A., Sandin, S., and Rhodes, D. (2008) Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc Natl Acad Sci U S A*. 105: 8872-8877.

[120] Wang, J.-P., Fondufe-Mittendorf, Y., Xi, L., Tsai, G.-F., Segal, E., and Widom, J. (2008) Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. *PLoS Comput Biol*. 4: e1000175.

[121] Staynov, D.Z., and Proykova, Y.G. (2008) Topological constraints on the possible structures of the 30 nm chromatin fibre. *Chromosoma*. 117: 67-76.

[122] Woodcock, C.L., Grigoryev, S.A., Horowitz, R.A., and Whitaker, N. (1993) A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proc Natl Acad Sci U S A*. 90: 9021-9025.

[123] Bustamante, C., Marko, J.F., Siggia, E.D., and Smith, S. (1994) Entropic elasticity of lambda-phage DNA. *Science*. 265: 1599-1600.

[124] Wong, H., Victor, J.-M., and Mozziconacci, J. (2007) An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLoS One*. 2: e877.

[125] Demarchi, F., D'Agaro, P., Falaschi, A., and Giacca, M. (1993) In vivo footprinting analysis of constitutive and inducible protein-DNA interactions at the long terminal repeat of human immunodeficiency virus type 1. *J Virol*. 67: 7450-7460.

[126] Cron, R.Q., Bartz, S.R., Clausell, A., Bort, S.J., Klebanoff, S.J., and Lewis, D.B. (2000) NFAT1 enhances HIV-1 gene expression in primary human CD4 T cells. *Clin Immunol*. 94: 179-191.

[127] Chan, J.K.L., and Greene, W.C. (2011) NF-κB/Rel: agonist and antagonist roles in HIV-1 latency. *Curr Opin HIV AIDS*. 6: 12-18.

[128] Romanchikova, N., Ivanova, V., Scheller, C., Jankevics, E., Jassoy, C., and Serfling, E. (2003) NFAT transcription factors control HIV-1 expression through a binding site downstream of TAR region. *Immunobiology*. 208: 361-365.

[129] Tripathy, M.K., Abbas, W., and Herbein, G. (2011) Epigenetic regulation of HIV-1 transcription. *Epigenomics*. 3: 487-502.

[130] Lusic, M., Marcello, A., Cereseto, A., and Giacca, M. (2003) Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter. *EMBO J*. 22: 6550-6561.

[131] Rafati, H., Parra, M., Hakre, S., Moshkin, Y., Verdin, E., and Mahmoudi, T. (2011) Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency. *PLoS Biol*. 9: e1001206.

[132] Mastrangelo, I.A., Courey, A.J., Wall, J.S., Jackson, S.P., and Hough, P.V. (1991) DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. *Proc Natl Acad Sci U S A*. 88: 5670-5674.

[133] Levitt, M. (2007) Growth of novel protein structural data. *Proc Natl Acad Sci U S A*. 104: 3183-3188.

PUBLICACIONES

En esta sección se adjuntan los trabajos publicados en el marco de la presente Tesis de Doctorado (ver Tabla 3.1).

TABLA 3.1 | Lista y descripción de artículos publicados. Los artículos se numeran de acurdo al orden de referencia en el texto de la tesis.

Artículo N°	Descripción
1	Dans PD, Zeida A, **Machado MR**, Pantano S. (**2010**) A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory. Comput.* 6:1711-1725
2	Zeida A, **Machado MR**, Dans PD, Pantano S (**2012**) Breathing, bubbling and bending: DNA flexibility from multi microseconds simulations. *Phys. Rev. E. 86: 021903* . DOI: 10.1103/PhysRevE.86.021903
3	Darré L, **Machado MR**, Dans PD, Herrera FE, Pantano S. (**2010**) Another coarse-grain model for aqueous solvation: WAT four? *J. Chem. Theory. Comput.* 6: 3793-3807
4	Dans PD, Darré L, **Machado MR**, Zeida A, Pantano S. (**2011**) Coarse grain potential: a model for DNA in implicit and explicit solvent. En “A course on biomolecular simulations”, ed. Jordi Villà-Freixa, Huygens Editorial, en prensa.
5	**Machado MR**, Dans PD, Pantano S. (**2011**) A hybrid all-atom/coarse grain model for multiscale simulations of DNA. *Phys. Chem. Chem. Phys.* 13: 18134-18144
6	**Machado MR**, Dans PD, Pantano S. (**2010**) Isoform-specific determinants in the HP1 binding to histone 3: insights from molecular simulations. *Amino Acids.* 5:1571-1581

JCTC Journal of Chemical Theory and Computation

A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics

Pablo D. Dans, Ari Zeida, Matías R. Machado, and Sergio Pantano*

Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay

Received December 4, 2009

Abstract: Coarse-grain (CG) techniques allow considerable extension of the accessible size and time scales in simulations of biological systems. Although many CG representations are available for the most common biomacromolecules, very few have been reported for nucleic acids. Here, we present a CG model for molecular dynamics simulations of DNA on the multi-microsecond time scale. Our model maps the complexity of each nucleotide onto six effective superatoms keeping the "chemical sense" of specific Watson−Crick recognition. Molecular interactions are evaluated using a classical Hamiltonian with explicit electrostatics calculated under the framework of the generalized Born approach. This CG representation is able to accurately reproduce experimental structures, breathing dynamics, and conformational transitions from the A to the B form in double helical fragments. The model achieves a good qualitative reproduction of temperature-driven melting and its dependence on size, ionic strength, and sequence specificity. Reconstruction of atomistic models from CG trajectories give remarkable agreement with structural, dynamic, and energetic features obtained from fully atomistic simulation, opening the possibility to acquire nearly atomic detail data from CG trajectories.

Introduction

Computer simulations have become a reliable tool for the study of structure and dynamics of soft condensed matter systems, as they expose molecular insights that can be difficult or impossible to obtain with experimental techniques. The continuous motivation to expand the limits imposed by the available computer power has prompted scientists to develop simplified representations that reduce the complexity, size, and conformational degrees of freedom of molecular systems while keeping the physical essence of the interactions that rule their behavior.[1] The remarkable improvement in accuracy and reliability achieved by the so-called coarse-grain (CG) representations, together with the development of new algorithms and computer power, offers currently the possibility to reach biologically relevant time scales and system sizes (see ref 2 for an exhaustive review of the latest developments in CG techniques applied to molecular systems). A wide variety of CG representations are available for the most common biological macromolecules, including highly complex lipid−protein systems (see, for instance refs 3 and 4). Nevertheless, only a few implementations have been reported for nucleic acids. Among these applications, notable success has been achieved in the description of DNA structure, dynamics, and melting.[5−8] At the base level, some interesting DNA models inspired us in developing our CG model. Zhang and Collins described the B-DNA as a sequence of rigid bodies (base-ribose) connected by flexible rods. Depending on the type of nucleic base (A/T or G/C), four to five centroids were used in the contraction scheme. Molecular dynamics simulations of thermal melting transition were performed using DNA fragments of 100 base pairs (bp).[9] Tepper and Voth developed a DNA model with explicit solvent particles using 14 uniformly distributed centroids per base pair, covalently linked to reproduce the spontaneous formation of the double helix.[5] In the model by Knotts et al.,[6] each base was reduced to three interaction sites with *ad hoc* potentials for stacking and base pairing. This model successfully reproduced salt-dependent melting, bubble formation, and rehybridization. Using wavelet projection to obtain the effective CG potential between effective centroids, the overall deformation response of a DNA

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

10.1021/ct900653p
Published on Web 04/14/2010

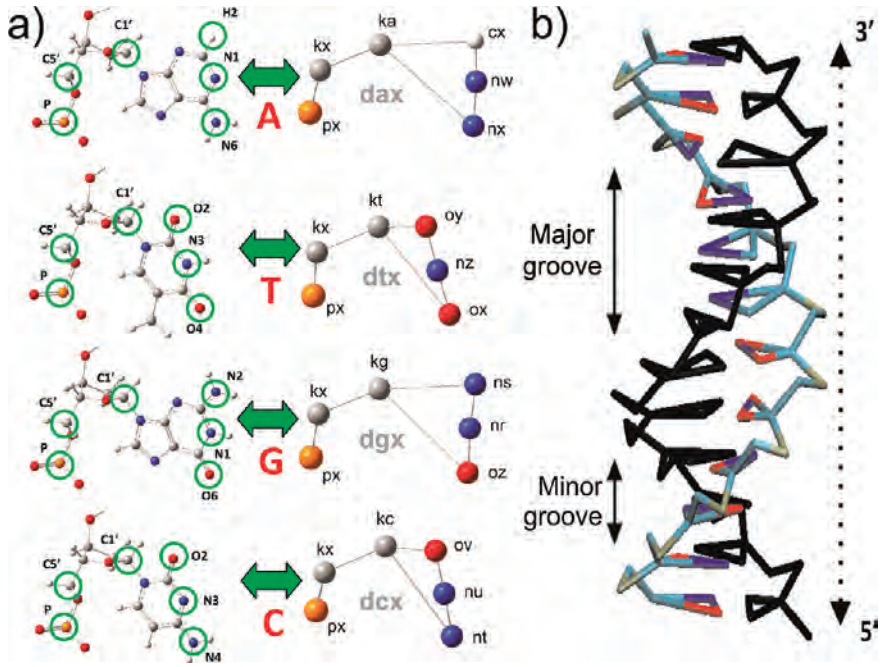

Figure 1. Mapping scheme between atomistic and CG models. (a) Circles highlight the coordinates of the elements from the all-atom representation preserved in the CG model. The residue, superatom, and connectivity are displayed. (b) CG representation of a 12-mer double helix DNA in the canonical B-form that illustrates grooves and 5′−3′ direction (black strand).

molecule was achieved with molecular dynamics (MD) techniques.[7] Representing the DNA as a worm-like polymer and using the "rigid base pair model", homogeneous elastic properties were reproduced by fitting the model against experimental data.[8] In the Mergell et al. model of DNA, each base pair was represented by a rigid ellipsoid linked to the backbone by semirigid harmonic springs.[10] Recently, CG models of DNA were devoted to protein−DNA docking, by optimizing the interaction surface between the macromolecular partners.[11] Similarly, simplified Go-models for RNA have accomplished the description of folding dynamics under varying temperatures and mechanical stretches.[12,13] With a less detailed representation, RNA[14] and also DNA[15] molecules were reduced to only one centroid per nucleotide to study the packing dynamics of a virus genome inside the protein capsid. In this last DNA study, an implicit solvent approach was used to mimic the biological environment.[15] These kinds of models have also been applied with success to the description of large molecular aggregates such as nucleosomes and ribozymes.[16−22]

In this contribution, we present a new CG model for MD simulation of nucleic acids ruled by a Hamiltonian function identical to that used by the most popular MD simulation packages. Electrostatic interactions are treated within the framework of the generalized Born model for implicit solvation.

The model reproduces canonical structures as well as conformational transitions from the A to B form of DNA. We obtain also a good reproduction of the temperature, size, and sequence-specific and ionic strength driven melting. The breathing dynamics of poly(AT) domains were compared with experiments raising comparable life times for end-fraying and also internal hydrogen bonds disruption at the base pair level. Reconstruction of all-atom trajectories from CG MD runs shows a high-quality reproduction of geometrical features with maximum deviations on the order of 2−3 Å with respect to the experimental structures and/or all-atom simulations.

Methods

Coarse Grain Mapping. Our CG model reduces the complexity of a nucleotide to six effective interaction sites (hereafter called superatoms) for each type of canonical nucleotide in DNA (A, T, C and G). This defines four different coarse-grained bases (dax, dtx, dcx, and dgx), which map to the all-atom nucleotides as illustrated in Figure 1a retaining the "chemical sense" of the interactions. Each of the six superatoms was placed on the Cartesian coordinates of one element in the all-atom representation and condensed the molecular information from its atomic neighborhood. The number of superatoms chosen retains the Watson−Crick interaction sites and preserves the asymmetry in the backbone, the identity of the minor and major grooves, as well as the 5′−3′ polarity of the DNA strands (see Figure 1b). Under this scheme, the total mass of the individual atoms of the real nucleotides, including hydrogen, is condensed onto the superatoms, as shown in Table 1.

Phosphate groups are represented by the px superatoms placed on the position of the corresponding phosphorus. The position of the C5′ atom was used to place the superatom kx, which serves to establish the 5′−3′ direction of each DNA strand and allows for the formation of the major and minor grooves (see Figure 1b). The kn superatom (where kn = ka, kt, kc, or kg) lays at the position of the C1′ atom. The superatoms that participate in the Watson−Crick interactions are placed in the same position as the corresponding atoms preserving the molecular specificity between

Table 1. Masses, Charges, and Lennard-Jones Parameters Assigned to the Superatoms

superatoms[a]	mass	atoms represented[b]	charges (e)	Lennard-Jones ε (kcal/mol)	Lennard-Jones σ (Å)
px	78.97	P+O1P+O2P+O5′	–1.00	0.2000	2.6000
kx	73.07	C5′+C4′+C3′+O3′+O4′	0.00	0.1094	2.4080
ka	41.05	C1′+C2′+N9	0.00	0.1094	1.9080
nx	40.03		0.35	0.1900	1.8240
nw	40.03	(C8+N7+C5+C4+C6+N6+N1+N3+C2)[c]	–0.35	0.1900	1.8240
cx	40.03		0.00	0.1094	1.9080
kt	41.05	C1′+C2′+N1	0.00	0.1094	1.9080
ox	37.03		–0.35	0.2400	1.6612
nz	37.03	(C6+C5+O4+C4+N3+O2+C2+C)[c]	0.70	0.1900	1.8240
oy	37.03		–0.35	0.2400	1.6612
kg	41.05	C1′+C2′+N9	0.00	0.1094	1.9080
oz	45.71		–0.70	0.3100	1.6612
nr	45.71	(C8+N7+C5+C4+N3+C2+N1+C6+O6+N2)[c]	0.35	0.2600	1.8240
ns	45.71		0.35	0.2600	1.8240
kc	41.05	C1′+C2′+N1	0.00	0.1094	1.9080
nt	32.03		0.70	0.2600	1.8240
nu	32.03	(C6+C5+C4+N4+N3+O2+C2)[c]	–0.35	0.2600	1.8240
ov	32.03		–0.35	0.3100	1.6612

[a] The types of the superatoms match those included in the coordinate and topology files that are available from the authors upon request. [b] Hydrogen atoms are omitted for brevity. Their masses are added to the corresponding heavy atoms. [c] The sum of the masses is equally distributed among the three superatoms.

both DNA strands. In this sense, all-atom Watson−Crick hydrogen bonds are shrunk to two-point electrostatic interactions in the CG model.

This scheme leads to an easy mapping/back-mapping from all-atom to CG representation and vice versa. Using internal coordinates and canonical distances, angles, and dihedrals from the B-form of Arnott et al.,[23] we can recover the complete all-atom picture. Dynamic events in the ps−ns time scale can be followed within a multi-microsecond trajectory calculated at the CG level. To this aim, we developed an algorithm that uses as input the instantaneous position of three superatoms to infer the Cartesian coordinates of the atoms in the neighborhood in each MD frame. A Fortran 90 implementation of the homemade algorithm is provided in Table S1 as Supporting Information. The reconstruction to the all-atom picture is made in three steps proceeding from the base to the phosphate moiety (see Table S2 in the Supporting Information for a pseudo-code explaining the algorithm). Since we have less information about the sugar conformation and the dihedrals involved in the phosphodiester bond, a loss of accuracy of the back-mapped coordinates in the backbone region can be expected (see Figure S1 in the Supporting Information). To correct the positioning of the sugar moiety and the distances of the phosphodiester bonds, 150 steps of geometric optimization were performed on each frame after the complete CG to all-atom reconstruction (see Figure S2 in the Supporting Information).

Parameterization. With the aim of maximizing the transferability between different MD packages, our model employs a widely used Hamiltonian function:

$$U = \sum_{\text{bonds}} k_b(r_{ij} - r_{eq})^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2}[1 + \cos(n_k\varphi - \gamma_k^{eq})] + \sum_{l}^{N}\sum_{l>m}^{N}\left\{4\varepsilon\left[\left(\frac{\sigma}{r_{lm}}\right)^{12} - \left(\frac{\sigma}{r_{lm}}\right)^{6}\right] + \frac{q_l q_m}{\in r_{lm}}\right\} \quad (1)$$

where k_b is the bond stretching constant, $r_{ij} = r_i - r_j$, and r_{eq} is the equilibrium bond distance between two linked elements. k_θ is the bond angle constant. θ is the instantaneous angular value defined by three successive elements, and θ_{eq} is the equilibrium bond angle. V_k is the height of the torsional barrier; n_k is its periodicity. φ is the torsion angle defined by four consecutively bonded elements, and γ_k^{eq} is the phase angle. In the fourth term, the sum runs over all the particles of the system (N). This term corresponds to the Lennard-Jones and Coulombic potentials, in which ε is the maximum depth of the function and σ is the zero energy point or van der Waals diameter. While the values of ε were used as free parameters, those of σ for the backbone superatoms were set to roughly match the excluded volume of the groups of atoms represented (see Table 1). Superatoms participating in the base preserve the σ values coming from the corresponding heavy atoms to avoid artifacts that could disrupt the intra-base-pair step (rise). Lastly, $q_{l,m}$ is the charge of each superatom, and $\in$ is the vacuum permittivity.

Hydration and ionic strength effects were taken into account using the generalized Born (GB) model[24] for implicit solvation as implemented in AMBER.[25] The Born effective radii were fixed to 1.5 Å for all superatoms.

In the present model, the equilibrium bond distances and bond angles were taken from the canonical B-form of Arnott et al.[23] The bond stretching and bond angle constants were fixed to 400 kcal/mol·Å^2 and 75 kcal/mol·rad^2 for all bonds and angles, respectively (eq 1). The torsional barrier for the three dihedral angles of the backbone was fixed to 10 kcal/mol (see Φ, Ξ, and Ψ in Figure 2 and Table 2). The periodicity of dihedral angles was set to nearly reproduce the canonical conformations of the B-form of Arnott et al.[23] To complete the model, two more torsionals, Γ_{dnx} and Ω_{dnx}, that act on the same bond as Ω were added (where dnx stands for each of the four bases: dax, dtx, dgx, and dcx). The parameters for the Γ_{dnx} and Ω_{dnx} dihedral angles, which can be visualized in Figure 2, are specific for each nucleic base. All the torsional parameters used in our model are displayed in Table 2.

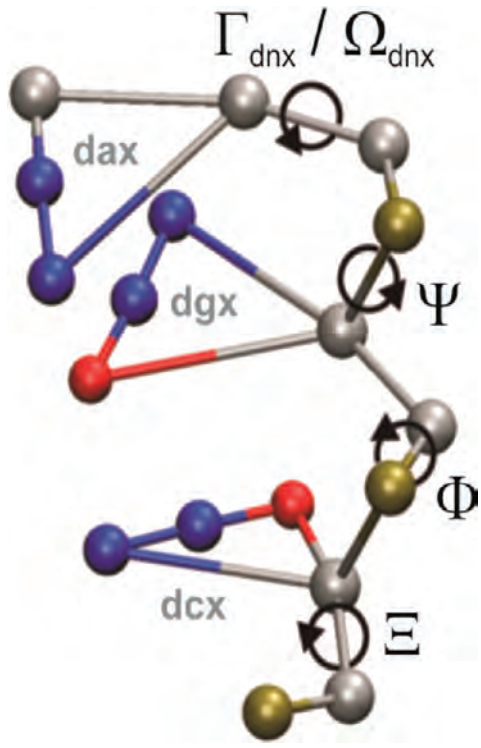

Figure 2. Dihedral angles used in the CG model. Three dihedrals account for the backbone movements for which the parameters are the same regardless of the nucleobase (Φ = kn-*px*-*kx*-kn, Ξ = px-*kx*-*kn*-px, and Ψ = kx-*kn*-*px*-kx where kn = ka, kt, kc, or kg). The dihedral angles Ξ, Ω_{dnx}, and Γ_{dnx} act on the same bond but are defined using different superatoms (dnx = dax, dtx, dgx, dcx). See Table 2 for dihedral angles definition.

Benchmark System: The Drew−Dickerson Dodecamer. To validate the structural, dynamical, and energetic behavior of our CG scheme, the results presented in the first part of this contribution correspond to the Drew−Dickerson dodecamer of DNA (also called the *Eco*RI dodecamer),[26−28] which was used as a benchmark system. This dodecamer of sequence 5′-d(CGCGAATTCGCG)-3′ has been largely studied by means of experimental and theoretical works, giving rise to a solid bibliographic base to compare our results.[29−33] As the starting structure for the CG simulation (labeled DDcgB), the Drew−Dickderson dodecamer was built[25] in the canonical B-form of Arnott el al.[23] During simulation, nonbonded interactions were calculated up to a cutoff of 18 Å within the GB approximation, and the salt concentration was set to 0.15 M. Temperature was controlled using a Langevin thermostat[34,35] with a friction constant of 50 ps^{-1}, which approximates the physical collision frequency for liquid water.[36] The random seed generator of the stochastic force was randomly changed every restart of the simulation (every 1 μs) to avoid quasi-periodic oscillations. The temperature was raised linearly from 0 to 298 K in 5 ns. After that point, production runs of 5 μs were performed, and snapshots were recorded for analysis every 50 ps using a time step of 5 fs to integrate the classical equation of motion. To avoid the fraying of the helix ends frequently observed in long MD simulations,[37] loose harmonic restraints of 3.0 kcal/mol·Å^2 were added to preserve the Watson−Crick hydrogen bonds of the capping base pairs.

To compare our results with state-of-the-art molecular dynamic simulations, the same sequence was built in the Arnott B-form,[23] solvated with explicit water molecules, and surrounded by K^+ and Cl^- ions to mimic the physiological conditions (this system was labeled DDaaB). The all-atom molecular dynamic simulation of the unconstrained Drew−Dickerson dodecamer was performed using the parm99[38] force-field with the correction proposed by Orozco and coworkers for nucleic acids (parmbsc0).[39] Ions were treated with the same force-field. The final system contained 36 K^+, 14 Cl^-, and 3926 TIP3P water molecules[40] in a truncated octahedral box. Initially, the water molecules and ions were relaxed by 1000 steps of energy minimization imposing harmonic restraints of 25 kcal/mol·Å^2 to DNA. Subsequently, four energy minimization runs were performed (with the same number of steps) where the restraints on DNA were gradually reduced from 20 to 5 kcal/mol·Å^2. All optimizations and equilibration MD simulations were performed using constant volume. Long-range interactions were treated using the PME approach[41] with a 12 Å direct space cutoff. The last optimized structure was taken as the starting point for the MD simulations. The entire system was then heated from 0 to 300 K during a 200 ps MD run with harmonic restraints of 5.0 kcal/mol·Å^2 imposed to DNA at a constant volume. Final temperature and a constant pressure of 1 atm were then reached by coupling the system to the Berendsen thermostat and barostat, respectivelly.[42] Fifty nanoseconds of production MD simulation were performed in the isobaric−isothermal ensemble. An integration time step of 2 fs was used, and all

Table 2. Torsional Parameters Used in eq 1 for the CG-DNA Model[a]

	torsional parameters											
dihedral	V_1[b]	V_2	V_3	V_4	n_1	n_2	n_3	n_4	γ_1^{eq}	γ_2^{eq}	γ_3^{eq}	γ_4^{eq}
kn[c]-px-kx-kn (Φ)[c]	10.0				8				161.0			
px-kx-kn-px (Ξ)	10.0				8				–153.2			
kx-kn-px-kx (Ψ)	10.0				4				–29.3			
px-kx-ka-nx (Ω_{dax})	10.0	6.0	7.0	10.0	1	7	2	1	118.0	47.0	20.0	–220.0
px-kx-ka-cx (Γ_{dax})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kt-ox (Ω_{dtx})	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	–140.0
px-kx-kt-oy (Γ_{dtx})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kg-oz (Ω_{dgx})	10.0	6.5	7.0	10.0	1	6	2	1	110.0	90.0	20.0	–220.0
px-kx-kg-oz (Γ_{dgx})	6.0	4.0	2.0		1	3	4		65.0	145.0	130.0	
px-kx-kc-nt (Ω_{dcx})	10.0	5.0	7.0	10.0	1	8	2	1	117.0	47.0	20.0	–140.0
px-kx-kc-ov (Γ_{dcx})	6.0	4.0	2.0		1	3	4		65.0	135.0	130.0	

[a] See Figure 2 for a comprehensive identification of the Φ, Ξ, Ψ, Ω_{dnx}, and Γ_{dnx} angles. [b] See third term in eq 1. [c] Where kn = ka, kt, kc, or kg.

bond lengths involving hydrogen atoms were restrained using the SHAKE algorithm.[43]

Using the *ptraj* utility of AMBER,[25] root mean square deviations (RMSD) were calculated on all the superatoms/atoms of each residue. The mobility of the bases relative to the backbone was evaluated by comparing atomic B-factors against experimental data. We calculated the quotient between the B-factors of the phosphate atom/superatom and the central heavy atom/superatom engaged in the Watson−Crick interaction (N1 for purines and N3 for pyrimidines). The CG trajectories were back-mapped to all-atom representation and, together with the state-of-the-art MD simulations, analyzed with the program Curves 5.1[44] to monitor the effects of thermal fluctuations upon the major determinants of the B-DNA molecular structure. Root mean square fluctuations (RMSF) and time evolution were calculated for selected helical parameters. The *anal* module of AMBER[25] was used to calculate the interaction energies between bases, strands, GC pairs, and AT pairs in terms of electrostatic and van der Waals contributions. When analyzing back-mapped trajectories, in all the cases, only a discontinuous 50-ns-long trajectory containing the final 10 ns of each microsecond was taken into account for shortness. For comparison purposes, calculated properties were also obtained for crystallographic and averaged NMR derived data (PDB structures 1BNA[45] and 2DAU,[46] respectively).

All MD simulations were carried out using the *sander* module of AMBER 10.[25] Molecular drawings were performed with VMD 1.8.6.[47]

DNA Melting. The CG model was tested to reproduce thermal melting for several systems analyzing the effect of variable length, GC content, and ionic strength of the medium. The sequences chosen were taken from the recently determined experimental work by Owczarzy and co-workers:[48]

(i) 5′-d(ATCGTCTGGA)-3′ (seq10)
(ii) 5′-d(TACTAACATTAACTA)-3′ (seq15a)
(iii) 5′-d(GCAGTGGATGTGAGA)-3′ (seq15b)
(iv) 5′-d(GCGTCGGTCCGGGCT)-3′ (seq15c)
(v) 5′-d(AGCTGCAGTGGATGTGAGAA)-3′ (seq20)

Separated runs were carried out for ionic strengths of 0.07, 0.12, 0.22, and 1.0 M. The melting protocol was the same for each sequence studied and consisted of 3.0 μs of MD simulation in which the temperature was raised 100 °C in five steps of 20 °C. Each step consisted of 0.1 μs of heating followed by 0.5 μs simulated at constant temperature. No restraints were added to the capping base pairs.

To define a melting criterion, hydrogen bonds between base pairs were considered to exist if the distance between the corresponding "acceptor" and "donor" superatom was less than 4.0 Å. The characteristic melting temperature is reached when 50% of the base pairs are in an open state. To generate the melting curves, the percentage of the opened base pairs within the sequence was calculated for each frame of the simulation. Adjacent averaging every 500 frames was performed to clean out the noise. Averaged points were sorted from lowest to highest temperatures, and a sigmoid fit with the Gompertz 4 parameters equation was applied:

$$y_0 + ae^{-e^{-(T-T_0)/b}} \quad (2)$$

This procedure yields one single continuous function of temperature. In eq 2, T_0 is the abscissa of the inflection point, which corresponds to the calculated melting temperature. The regression coefficients for all the sigmoid fits were always >0.8. Results were integrally obtained from the total CG trajectories. Notice that the back-mapping procedure was not applied.

The A to B Transition. The Drew−Dickerson sequence was also built in the A-form of Arnott et al.[23] to test the capability of the model to reproduce a conformational transition from the A to the B form (DDcgA). Five microseconds of coarse grained MD simulations were run under the same conditions used in the DDcgB system. RMSDs with respect to the experimental and canonical B-form structures, pitch, and minor and major groove width were calculated to evaluate the structural transition.

DNA Breathing Dynamics. Finally, we studied the breathing movement of the Drew−Dickerson dodecamer and a 29-bp-long double-stranded DNA: 5′-d(GGCGCCCAATATAAAATATTAAAATGCGC)-3′. The sequence contains a GC clamp domain (G1 to C7) and a long AT track that corresponds to a breathing domain (A8 to A24). The simulation conditions were fixed to roughly match the experimental work by Altan-Bonnet and co-workers.[49] The most relevant difference resided in the fact that the sequence used by Altan-Bonnet et al. contained a thymine tetraloop to avoid the separation of both strands. However, since the structure of this loop is unknown, we decided to replace it by loose harmonic restraints of 3.0 kcal/mol·Å^2 to preserve the Watson−Crick hydrogen bonds of the last base pair (5′-C_{29}-3′ in strand1 and 5′-G_1-3′ in strand2).

The criterion to define the base opening/closing was identical to that established for melting. MD simulations of 4 μs at 37 °C with an ionic strength of 0.1 M were performed.

Results and Discussion

A major goal for molecular simulations is not only the reproduction of stable trajectories of molecular systems oscillating around equilibrium conformations but also to achieve the capacity to explore the accessible conformational space and evolve toward more stable conformations. In the following paragraphs, we provide some examples of the performance of our model to reproduce the structure, energetics and dynamics of stable trajectories around equilibrium configurations, melting of DNA, conformational transitions, and breathing dynamics.

Benchmark System: CG Model vs All-Atom. All simulations started with the canonical B-form and were stable along all the simulation time. A first measure of the quality of the CG model can be obtained from a direct comparison between the whole trajectories of CG and all-atom representations. To this aim, we calculated the RMSD using all the superatoms in the CG model and the corresponding atoms in the all-atom trajectory (according to the mapping presented in Figure 1). We found that the intrinsic fluctuations during CG and all-atom schemes were very similar. Furthermore, the structural models obtained from both simulations with respect to the experimental structures are practically identical

Table 3. Structural Comparison between CG and All-Atom Simulations[a]

	mean during MD trajectory	starting conformer (B form)	X-ray (1BNA[45])	NMR (2DAU[46])
DDcgB	1.0 ± 0.3	1.8 ± 0.3	2.3 ± 0.3	3.1 ± 0.3
DDaaB	1.6 ± 0.4	2.8 ± 0.4	2.6 ± 0.4	2.7 ± 0.4

[a] RMSD are calculated over 5 μs and 50 ns for the CG and all-atom trajectories, respectively. Values are reported in Angstroms.

(Table 3). Only subtle differences appear when comparing both trajectories against the reference structures.

To analyze the internal flexibility of the dodecamer, B-factors were calculated for selected groups of atoms/superatoms and were compared with the values coming from the X-ray experiments (PDB structure 1BNA). Absolute B-factors calculated from the all-atom trajectory differ significantly from those determined using the CG approach and the X-ray experiments. Only global qualitative trends for the structure as a whole could be obtained. However, the B-factors of the phosphorus elements relative to those of atoms belonging to the base moiety are good descriptors of the relative mobility of different segments of the nucleobases. A comparison between these values indicates that the all-atom simulation (DDaaB) always has the highest mobility, while the coarse-grained version (DDcgB) always presents the lowest (Figure 3). As shown, the relative values were always greater than 1.0 for all the systems, pointing out, as expected, the higher mobility of the backbone with respect to the base. In general, we observe that the relative mobility is lower in the CG model. This can be related to the reduced number of degrees of freedom or to a nonoptimal mass distribution.

Benchmark System: Back-Mapped CG Model vs All-Atom. Despite these encouraging results, it becomes difficult to establish a direct comparison between both simulations. Therefore, we sought to extract atomistic information from our CG model. To this end, we back-mapped the last 10 ns of each microsecond from our CG trajectory (DDcgB). This generated an atomistic noncontiguous 50-ns-long trajectory that is directly comparable with that of the all-atom simulation (DDaaB).

Table 4. Structural Comparisons for the Drew−Dickerson Structure d(CGCGAATTCGCG)$_2$[a,b]

	DDcgB	DDaaB	Arnott-A	Arnott-B	1BNA	2DAU
DDcgB			6.5	1.8	2.3	3.1
DDaaB			5.6	3.0	2.8	2.8
Arnott-A	1.7	2.0		6.3	6.0	4.8
Arnott-B	0.9	1.5	1.5		1.4	3.4
1BNA	1.3	1.2	1.9	0.9		3.3
2DAU	1.5	1.4	1.9	1.5	1.6	

[a] Heavy-atom RMSD between the specified structures. [b] The upper-right portion represents RMSD fit measured in Å calculated over all the atoms. The lower-left portion represents RMSD fit calculated for the four base pairs underlined in the heading, i.e., residues 3−6 and 19−22.

Structural and Dynamical Comparison. Table 4 presents a comparative view of both simulations against the canonical A and B conformations and two experimental structures. The averaged RMSD for the DDaaB simulation was 2.8 Å apart from both the crystallographic (1BNA) and NMR (2DAU) structures and 3.0 Å with respect to the canonical B-form. Analogously, the family of structures obtained with the CG model remained 2.3 Å, 3.1 Å, and 1.8 Å apart from the X-ray structure, NMR structure, and canonical B-form, respectively (upper-right portion of Table 4).

If we consider the averaged RMSD calculated for the selected inner four base pairs (residues 3−6 and 19−22), the values are almost the same between DDcgB and DDaaB with respect to both experimental structures (lower-left portion in Table 4). We can conclude that the differences between all-atom and back-mapped CG simulations are rather subtle, and that both simulations sample very similar or equivalent conformational spaces.

A more stringent evaluation of the quality of the B-form reached by the CG model can be obtained from a comparison of the fluctuations of some selected helical parameters (Figure 4). RMSFs were calculated for the Slide, Rise, Roll and Twist, which are the most distinctive base pairs parameters between the A and B canonical forms (Figure 4a). The large fluctuations observed in the helix ends of DDaaB were not present in DDcgB due to the loose harmonic restraints imposed to preserve the Watson−Crick hydrogen bonds of the capping base pairs in the implicit solvent simulation.

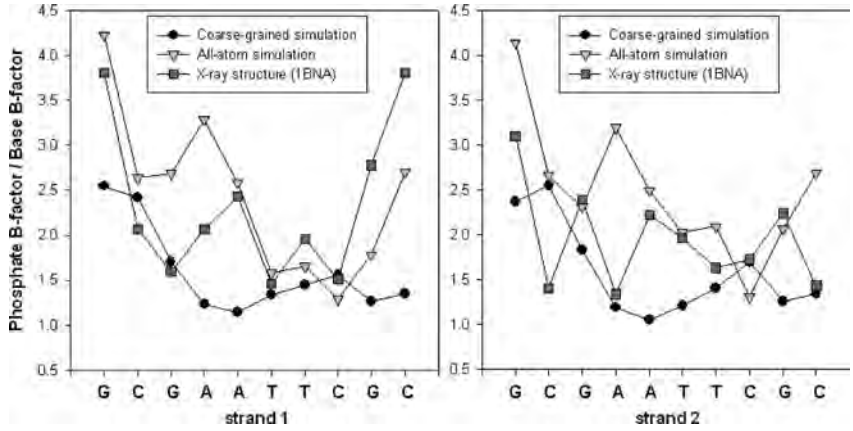

Figure 3. Higher mobility of phosphate groups. B-factors for the phosphorus atoms/superatoms relative to the central elements in the Watson−Crick interaction region along both strands. The coarse-grained (DDcgB) and the all-atom simulation (DDaaB) were compared to the experimental B-factors obtained from the X-ray structure with the PDB code 1BNA.

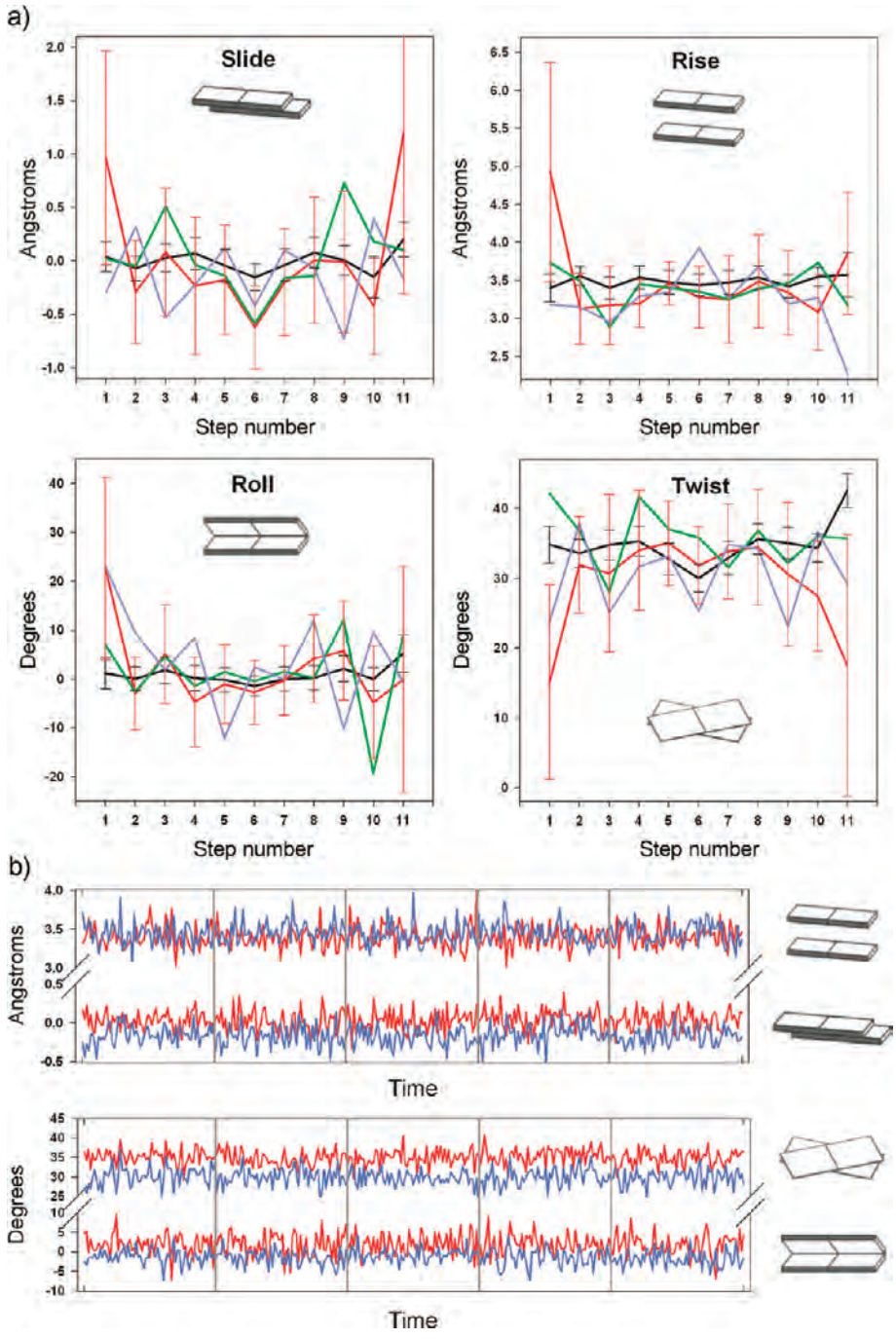

Figure 4. Selected helical parameters. (a) RMSF of the Slide, Rise, Roll, and Twist. The red line corresponds to DDaaB and the black line to DDcgB. Experimental structures 1BNA and 2DAU are represented by the green and the blue lines, respectively. Average values and standard deviations are plotted in Angstroms for the Slide and Rise and in degrees for the Roll and Twist parameters. The values are presented along the helix from the 5′ to 3′ direction (*x* axis). (b) The same helical parameters for two selected intra-base steps (C3/G4 in blue and A6/T7 in red) were plotted along 50 noncontiguous nanoseconds of the back-mapped DDcgB simulation.

Although the fluctuations about the mean values were in general somewhat larger in DDaaB versus DDcgB, the averages exhibited similar trends, especially in the Slide and Twist parameters. Compared to the all-atom simulation, the coarse-grained model exhibited a similar sequence-dependent trend in the Slide and Twist parameters for the CG, GA, AA, AT, TT, and TC dinucleotides (DNA steps 3−8 in Figure 4a).

A more dynamical picture of the structural stability can be acquired following the instantaneous values of the helical parameters during the simulation time. The same selected helical parameters are plotted against time for the back-mapped noncontiguous 50 ns trajectory. For the sake of brevity and clarity, only the C3/G4 and A6/T7 dinucleotides are plotted in Figure 4b. A first global inspection of Figure 4b illustrates the stability of the simulation, as no drift could be observed in the values of the parameters against the simulation time. The Rise and Slide fluctuated around the canonical values, and the Roll showed a distinctive behavior between the C3/G4 and A6/T7 dinucleotides comparable with

Table 5. Comparison of Averaged[a] Electrostatic and van der Waals (VdW) Interactions

	electrostatic (kcal/mol)		VdW (kcal/mol)	
	DDcgB	DDaaB	DDcgB	DDaaB
St1[b] vs St2	1449 ± 38	1434 ± 68	–66 ± 5	–72 ± 8
G4-C21 bp	4 ± 3	6 ± 2	–2 ± 1	–1 ± 1
A5-T20 bp	19 ± 2	11 ± 2	–2 ± 1	–1 ± 1

[a] The averages were calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds. [b] St1 stands for strand 1 and St2 for strand 2.

that observed in the all-atom MD simulation.[50] The slight separation between the Twist and Roll traces observed in Figure 4b may suggest a sequence-specific behavior. To shed light on this issue, an exhaustive and systematic study of the helical parameters for all the possible unique combinations of dinucleotides and tetranucleotides (for a total of 146 possible combinations) should be carried out and compared against recent results coming from molecular dynamic simulations.[50,51] Such study is clearly beyond the scope of the present contribution.

Energetic Comparison. In order to further validate the back-mapping procedure and obtain further support on the equivalence between the conformational spaces sampled by the CG and atomistic models, we compared the nonbonded interaction terms of the energy. Calculations were done averaging the results in vacuum using in both cases the same force field (parm99) applied to the all-atom MD and back-mapped trajectories. Comparisons for the van der Waals (VdW) and electrostatic components of the interaction energy between (i) the two strands, (ii) the bases of a GC pair, and (iii) the bases of an AT pair are shown in Table 5.

In light of the correspondent values within the standard deviations, the electrostatic and VdW interactions between DNA strands were virtually the same for both simulations. The good correspondence between both nonbonded interaction terms points out that the conformational space sampled by the CG model was energetically compatible with the state-of-the-art molecular dynamics. Note that the electrostatic contributions in Table 5 are always positive numbers since we computed the Coulombic interaction between two negatively charged strands. When comparing selected GC or AT base pairs, some subtle differences in the averaged electrostatics arise between both approaches. In our back-mapped CG model, the GC base pairs are slightly more stable, whereas the AT base pairs showed an opposite trend. Aimed at acquiring a more global picture, we looked at the electrostatic interactions per residue. For this task, we computed a 12 × 12 electrostatic interactions matrix. The results are presented as an interaction map in Figure 5. A very good correlation between both maps can be observed, providing further support for the compatibility between both approaches.

DNA Melting. Experimentally, the melting temperature (T_0) can be defined for an ensemble of double-stranded DNA molecules as the temperature at which half of the population is in the double-helical state and half in "random-coil" states. This type of definition, which is a good approximation for short DNA sequences, matches with the assumption that

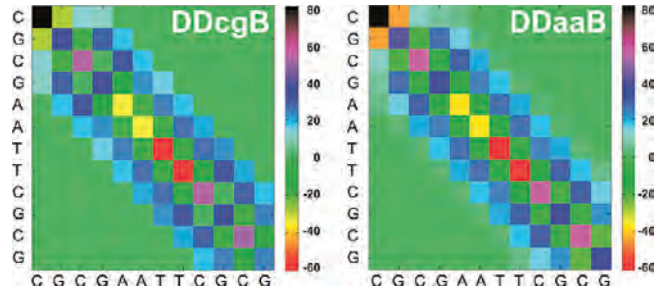

Figure 5. Color map of the averaged electrostatic interaction between the 12 nucleotides within the same strand. Comparison between the back-mapped coarse grained (DDcgB) and the all-atom (DDaaB) simulations. The color scale ranges from −60 to +80 kcal/mol, which are the lower and upper boundary values in the all-atom simulation. It must be noticed that these values were obtained from an effective force field and must not be taken as real energies. The average was calculated over 50 contiguous (DDaaB) or noncontiguous (DDcgB) nanoseconds.

melting occurs in a two-state transition. The melting temperature is highly dependent on the length of the double-stranded DNA. Furthermore, because GC base-pairing is generally stronger than AT base-pairing, the amount of guanine and cytosine (called the "GC content") can be estimated by measuring the temperature at which DNA melts. T_0 also depends on the salt concentration or ionic strength of the surrounding medium, as a higher electrostatic screening reduces the mutual repulsion between the negatively charged backbones of each strand in the macromolecule. In other words, T_0 can be used as an indirect measurement of the thermodynamic stability of a double-stranded DNA filament. In terms of the modeling, a good reproduction of the melting process may be indicative of a well-balanced energetic representation of the molecule under study.

To analyze the energetic features of the CG model, we followed the melting process of five sequences of different lengths, varying also the GC content and the ionic strength according to the Debye−Hückel screening parameter κ.[52] Our results were compared with recent experimental determinations for the same DNA sequences under nearly the same conditions.[48] No back-mapping was performed, as the fraction of native contacts can be measured directly from the CG trajectories.

We studied the length and GC-content dependence of the melting behavior for double-stranded DNA in implicit solvation. Melting temperatures were obtained from single simulations of double-stranded DNA where the temperature was raised in discrete steps of 20° to determine the melting point.

At first glance, good qualitative agreement can be found. As expected, increasing the base pairs number produced a higher T_0 (Figure 6a). Similarly, a higher GC content shifts the T_0 to higher temperatures (Figure 6b). However, in light of standard deviations in the temperature measurement (Table 6), the results could be considered rather qualitative.

There was no variation in T_0 for seq15b at 0.07, 0.12, and 0.22 salt concentrations, for which the calculated melting point was always 63 °C (see Table 6). The only significantly

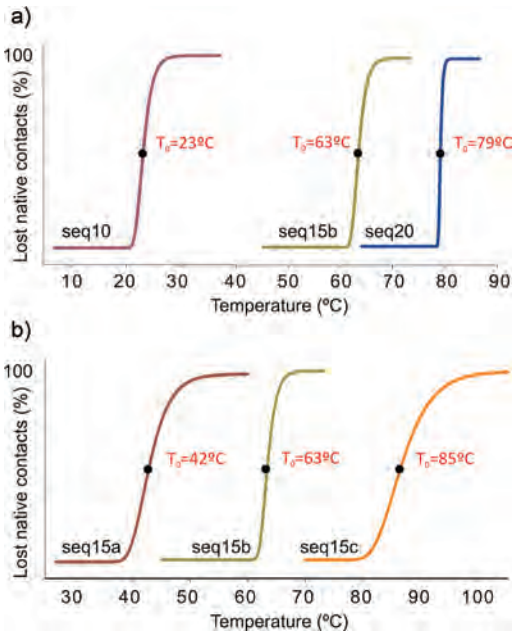

Figure 6. Fittted melting curves. (a) Sequences containing 10 (seq10), 15 (seq15b), and 20 (seq20) bp and 50−53% GC content. (b) Sequences with 15 bp for which the GC content is 20% (seq15a), 53% (seq15b), and 80% (seq15c), respectively. The inflection points (see eq 2) that determine the melting temperatures are indicated with black dots. Notice that the melting curves were obtained after a fitting procedure (see Methods). The numeric values along with the corresponding standard deviations are displayed in Table 6.

different T_0 was obtained at a 1.0 M salt concentration. This is probably due to the way in which the salt effects are incorporated into the GB model. In practice, the linearized Debye−Hückel approximation gives salt effects that are somewhat larger than those predicted by more accurate methods.[52] Saturation of salt effects takes place near 1.0 M, and the best fit with more accurate Poisson−Boltzmann estimations occurs for values from 0.1 to 0.4 M.[52] Previous MD simulations of nucleic acid structures carried out with either a 0.1 or 0.2 M salt concentration showed almost identical results.[53] Recent work describing the melting reaction in DNA hexamers using the same force field (parm99 with the Perez and co-workers modification[39]) and more accurate all-atom simulations for sampling of the free energy landscape also gave only qualitative results.[54]

The aim of this last set of simulations discussed was to test the qualitative dependence of the melting point upon variations of different factors. A precise determination of the melting temperature would need a better sampling such as, for instance, that performed by Knotts and co-workers.[6] They used replica exchange methods to achieve a more quantitative determination. We decided to not perform this kind of calculation, as there is a rather large arbitrariness in the molecular level definition of the melting point. For instance, a small variation (even of tenths of an angstrom) in the cutoff criteria for a native contact between two interacting bases can significantly shift the position of the melting points.

A clear advantage of using MD simulations is that the dynamic behavior of the melting process can be followed on the molecular scale. Thus, sequence- and location-dependent initiation and propagation of the steps that leads to DNA denaturation can be analyzed in detail. In all the sequences studied here, the melting of the helix started from the termini and proceeded toward the center (as an example, the movie for seq15b at 0.12 M is provided in the Supporting Information). This suggests that the loss of internal Watson−Crick interactions has a high-energy cost if the terminal base pairs are still formed as observed in other all-atom simulation work,[54] making internal fraying less frequent.

The A to B Transition. A celebrated result of effective force fields was the capability to reproduce complex conformational changes such as the A to B transition in duplex DNA.[55,56] Therefore, we faced the challenge of reproducing with our CG model the transition from the canonical A to B form, which is the physiologically more stable conformation of double-stranded DNA.

We prepared the same Drew−Dickerson dodecamer studied in the previous section but in the canonical A-form. To follow the A→B transition along the simulation, we calculated the RMSD of all the superatoms with respect to the corresponding atoms in the canonical B-form (see mapping scheme in Figure 1) and the two experimental structures. The results for 5 μs of simulation are shown in Figure 7. The conformational transition took place progressively in a relatively long time window, arriving at final state after nearly 1.2 μs (Figure 7a).

The final RMSD value reached after the transition was 3.3 Å with respect to the canonical B-form, e.g., a value comparable with the deviations obtained from atomistic simulations of duplex B-DNA using the generalized Born approximation.[37]

To reach the final B-form structure (between 1.2 and 5 μs), the conformational transition occurred in three steps:

Table 6. Reference Names and DNA Sequences Used in the Melting Experiments for which the GC Content and Salt Concentrations Are Indicated

reference name	DNA sequence (5′–3′)	GC content (%)	salt concentration (M)	T_0 exptl[a] (°C)	T_0 calcd (°C)	st. dev.
seq10	ATCGTCTGGA	50	0.12	37.4	23	25
seq15a	TACTAACATTAACTA	20	0.12	40.4	42	20
seq15b	GCAGTGGATGTGAGA	53	0.07	51.2	63	22
			0.12	54.8	63	22
			0.22	58.0	63	22
			1.00	63.3	100	26
seq15c	GCGTCGGTCCGGGCT	80	0.12	67.7	85	25
seq20	AGCTGCAGTGGATGTGAGAA	50	0.12	63.5	79	19

[a] Taken from ref 48.

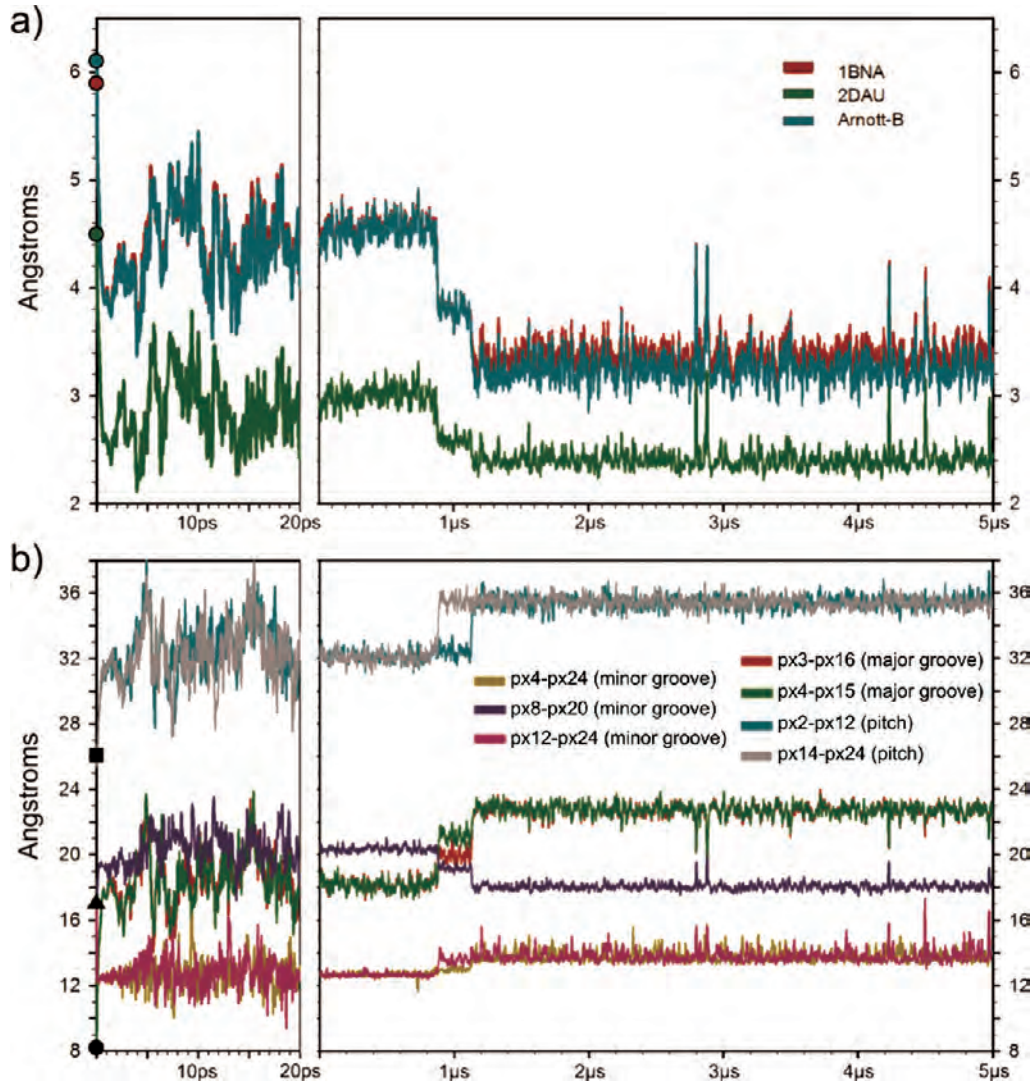

Figure 7. Time evolution of the A to B conformational transition. (a) RMSD using as a reference the canonical B-form (Arnott-B, blue line) and the X-ray and NMR structures (1BNA, dark red line, and 2DAU, green line, respectively). Colored dots indicate the RMSD of the initial conformer with respect to the reference structures. (b) Time evolution of selected distances (pitch, minor and major grooves) during simulation (color codes are indicated in the picture). Black squares, triangles, and circles indicate the starting values for pitch and minor and major grooves, respectively. In both cases, the data shown in the left panels correspond to instantaneous values, while data presented in the right panels correspond to a running average every 200 frames.

(i) In the first few picoseconds (left panel in Figure 7a), the initial structure (canonical A-form) underwent an abrupt conformational change that mainly affected the width of the major groove and, in a second degree, the overall pitch (see Figure 7b). On average, the major groove went from 8 to 18 Å and the pitch from 26 to 32 Å. These changes gave rise to a first cluster of structures 2.6 Å apart from the canonical A-form that remained stable during the first ~900 ns (step 1 in Figure 8). Using the generalized Born model, Tsui and Case[53] showed the convergence from an A-form DNA to a cluster of structures near the B-form within 20 ps of simulation. The quick transition was characterized by the rapid increase of the major groove and the end-to-end length (pitch). The same behavior was observed in the first 20 ps of the CG simulation (Figure 7b). Obviating that the DNA sequence is not strictly the same, visual inspection of the final structure obtained by Tsui and Case[53] after the transition looks very similar to the first cluster of structures obtained in the first picosecond of our CG model (compare the second structure in Figure 8 with Figure 9 in ref 53).

(ii) The following ~300 ns were characterized by a second cluster of structures 3.3 Å apart from the initial structure (first shoulder in Figure 7a). As shown in Figure 7b, the major groove continued to increase from 18 to 21 Å. This movement was followed by a decrease in the wideness of the minor groove measured in the central part of the sequence (from residues 8 and 20, dark blue line). In this case, the pitch underwent an asymmetric transformation to first rearrange the 3′−5′ strand; subsequently the 5′−3′ strand changed its value from 32 to 35 Å (a value very near the 34 Å of the canonical B-form).

(iii) Finally, between 1.2 μs and the end of the simulation, a last cluster of conformers 3.0 Å apart from the reference structure could be found. To reach this last state, the pitch in the 5′−3′ strand went to a final value of 35 Å. The major groove experienced a subsequent increase accompanied by

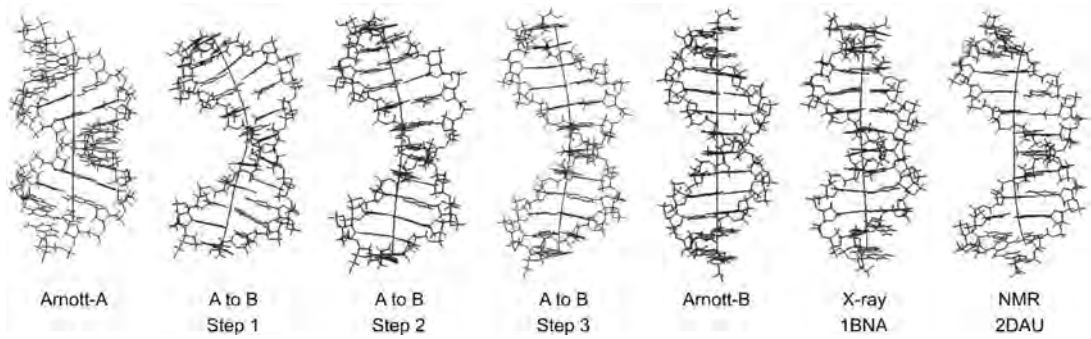

Figure 8. Comparison between back-mapped snapshots and atomistic structures. The conformers labeled steps 1−3 correspond to back-mapped representative snapshots from the conformational A to B transition: steps 1 (0−900 ns), 2 (900−1200 ns), and 3 (1.2−5.0 μs). The DNA axis was calculated with the Curves program.[44]

a ∼1 Å narrowing in the minor groove. Note that, along the 5 μs of simulation, the minor groove measured in the extremity of the sequence (between residues 4−24 and 12−16) only underwent slight changes.

In short, the A→B transition can be characterized by global changes in the major structural determinants of double-helical DNA (pitch and groove measurement) in a way that reminds the motion of a "crankshaft". Worth notice is the presence of some peaks in the RMSD after 2 μs of simulation. These correspond to little shifts between the two strands in the AT track that produce transient changes in the minor and major grooves. This behavior was only observed in the central tract and can be associated with breathing movements in the double helix (see next section).

As shown in Figure 8, the conformational changes seem to begin in the central part of the double helix and propagate to the ends, in the same way reported by Cheatham and Kollman in the first simulation on the A to B transition of DNA using all-atom simulations in explicit solvent.[53,55]

The comparison of the A to B transition with the work of Tsui and Case[53] appears to be relevant in the context of the actual time scale sampled by our CG scheme. This is always a complicated issue when dealing with CG simulations, as it is expected that the reduction of degrees of freedom translates to a flattening of the conformational space. The putative correspondence between our work and that of Tsui and Case seems to suggest some equivalence between both simulation schemes. However, the correspondence in the conformational transition may be an artifact of the model that is parametrized to reproduce the B-DNA. To further explore this issue, we sought to test our model against experimental data for which characteristic times ranging from picoseconds to hundreds of microseconds have been reported.

DNA Breathing Dynamics. The microsecond time scale for the full A to B transition begs the question of the correspondence between the real and simulated times. Some insights about this issue can be obtained from a comparison with published simulations on the microsecond time scale. Along the CG simulations of the Drew−Dickerson dodecamer, some transient base pair opening events occurred during the trajectory, especially at the AT pairs. The average lifetime of an open base pair is typically on the order of few picoseconds, but some opening events last for hundreds of picoseconds. These results are in very good agreement with the work of Perez and co-workers,[57] who performed the atomistic simulation of the Drew−Dickerson over 1.2 μs.

Aimed at directly comparing our model with well established experimental results and acquiring a more global perspective, we sought to perform the simulation of a 29-bp-long double-stranded DNA trying to mimic the laboratory conditions.[49] Base pair opening/closing dynamics have been reported for this kind of system on time scales ranging from picoseconds to nanoseconds[58] to hundreds of microseconds.[49] This would allow us to set the time frame of our simulations within a time scale window of near 8 orders of magnitude, covering (i) end-fraying, (ii) breathing, i.e, opening/closing of internal base pairs, and (iii) bubble formation, i.e., temporary opening of internal base pairs implying a partial loss of the double-helical structure.

Following the criterion to define an open state (see the Methods), we calculated the instantaneous state of each base pair (open/close) for each frame of the simulation and the time and sequence extension of those events. As was expected, significantly fewer open states were found in the GC clamp region compared to the AT domain (Figure 9a). Fraying events typically involved few base pairs (typically one or two, Figure 9b) that relax reaching the closed state in dozens to hundreds of picoseconds. This effect is compatible with X-ray,[59] NMR,[60,61] and computer[31,61] studies indicating that fraying is largely confined to the last two base pairs. The CG model also agrees with time-resolved Stokes shifts spectroscopy measurements that restrict the base-opening time to the range of dozens of picoseconds to a few nanoseconds.[62] Nevertheless, during the 4 μs of simulation, we found two events where the end-fraying spread even up to the sixth base pair (Figure 9b,c).

In the AT domain, a nearly continuous breathing dynamic was found along the simulation (Figure 9a), registering several opening/closing events. These events remained in the open state on the nanosecond time scale (see Figure 9b right). The global deformation and the time scale are well comparable with the NMR imino-proton exchange measurements.[58] In this technique, only slight opening of the base pairs, as those observed in the CG model, would be sufficient for the reaction to occur.

Notably, simultaneous opening/closing events with extensions from 2 to 10 consecutive base pairs were frequently observed (Figure 9b). Although with a much shorter time

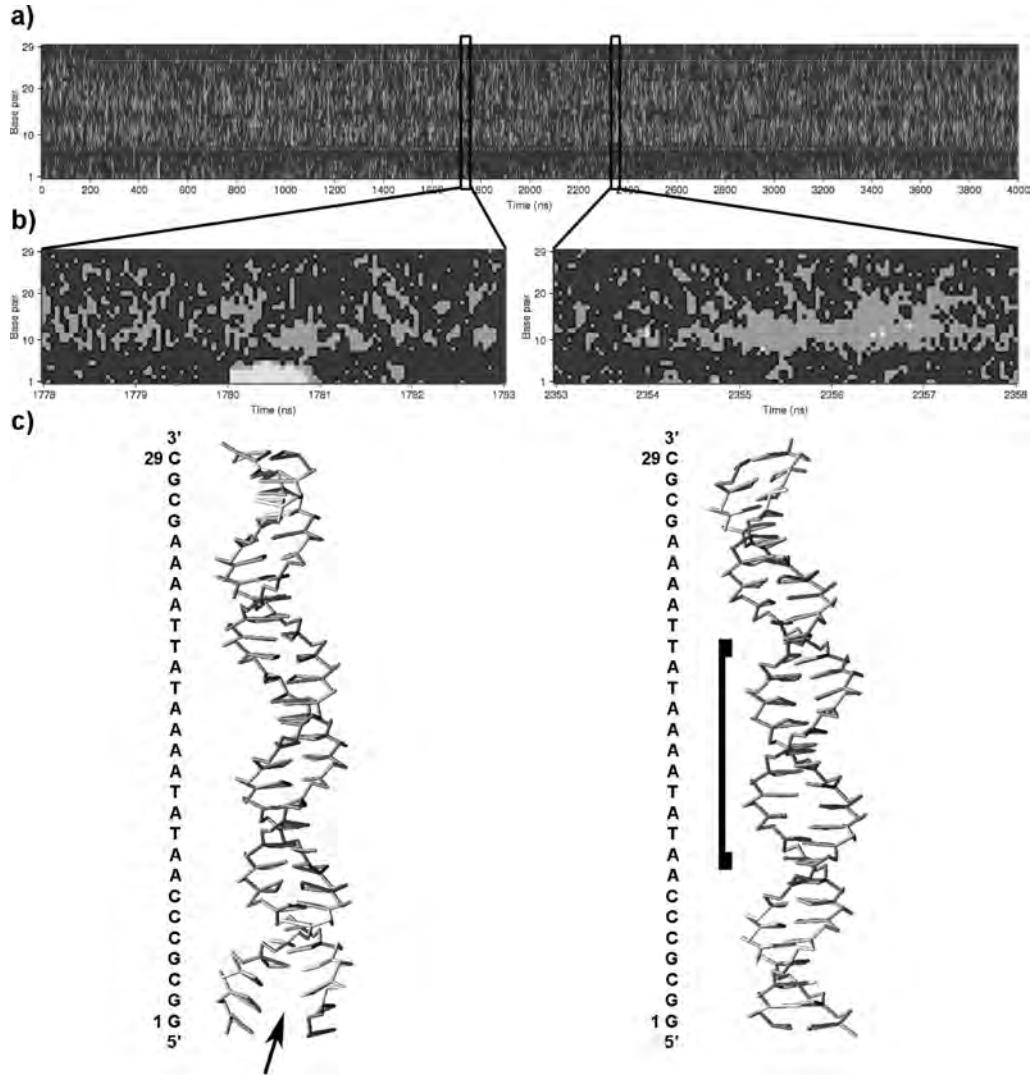

Figure 9. Breathing dynamics of the 29-bp-long double-stranded DNA. Base pairs (*y* axis) are plotted versus time (*x* axis) in nanoseconds. (a) Overview of the breathing along the trajectory. Dark gray color represents closed state base pairs (inter base distance lower than 4 Å). Open states were divided into two ranges: from 4 to 6 Å (light gray) and more than 6 Å (white). White dashed lines delimit the AT breathing domain.[49] (b) Five nanosecond closeups of the trajectory. (c) Representative structures of the end-fraying at the GC clamp (left) and AT breathing domain (right). Fraying and breathing are evidenced with an arrow and square bracket, respectively.

range, these results agree with multiexponential kinetics inferred from fluorescence relaxation times in an analogous molecular system for which opening/closing times of 20−100 μs were reported.[49] It is worth note that these data were obtained from fluorescence quenching experiments, which require a significant distortion in the double-helical structure (bubble formation) in order to be detectable. Such large deformations were never observed along our simulations.

The correspondence with previous theoretical work[57] and NMR studies[58] suggests that the time scale sampled by our model may roughly match the real one. Should this be true, a simulation time on the order of milliseconds would be needed to properly sample the ~100 μs process of bubble formation reported for 29-bp-long double-stranded DNA.[49] Alternatively, the absence of large deformations in our CG simulations could be related to the relative stiffness in the torsional parameters used. A larger number of simulations on different systems and comparison against experimental data are needed to further clarify this point.

Conclusions

We presented herein a nontopological CG model for MD simulations of DNA with explicit electrostatics that offers the possibility to fully recover the atomistic information. Back-mapped CG trajectories gave geometries with maximum deviations of a few angstroms from experimental values, which may be compatible with all-atom simulations offering a considerable speedup. Coarse-grained simulations were carried out in a single node with eight Intel Xeon 2.66 GHz cores at a rate of ~100 μs/superatom/day. At this rate, we performed 1 μs of the coarse-grained simulation using the Drew−Dickerson system in ~1.5 days. Around 850 days would be needed to run 1 μs of the all-atom simulation described herein. Globally, a speedup by a factor of nearly

600 is granted using the CG model. An advantage of the present contribution is that many of the published CG simulation schemes are implemented in *ad hoc* codes or require tailor-made modifications of standard simulation packages, which are often difficult to access and/or operate for the general public. A notable exception of this is the MARTINI force field.[63] The evaluation of the interactions using a classical Hamiltonian allows for a straightforward porting to any other publicly available MD simulation package (topologies and parameters files in AMBER format are available from the authors upon request).

Although the sampling time remains a not completely solved issue, this kind of implementation may open new alternatives to the study of dynamic properties of nucleic acids at longer time scales and for larger systems.

Finally, we would like to stress the fact that the results showed here cover only applications where DNA exists near its B-form. Clearly, Hoogsteen and sugar-edge pairs are out of reach for the present model. This begs the question of whether noncanonical structural motifs can be also well described (structure of telomeres, circular DNA, etc.). This is particularly relevant for the case of RNA where a multiplicity of structural motifs is present (bulges, wobbles, hairpins, and internal loops, etc.). Work is currently ongoing in our group to expand the description to these more challenging cases.

Acknowledgment. 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), and Grant FCE_60-2007. M.R.M. is a beneficiary of the National Fellowship System of ANII.

Supporting Information Available: Fortran 90 implementation of the homemade algorithm needed for the reconstruction of the CG trajectories is provided. A pseudocode version explaining the homemade algorithm and two figures illustrating its accuracy (before and after the energy minimization) are also provided along with a movie of the melting process for seq15b at a 0.12 M salt concentration. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321* (5890), 798–800.

(2) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Taylor & Francis Group: New York, 2009; pp 1−455.

(3) Treptow, W.; Marrink, S. J.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112* (11), 3277–3282.

(4) Ollila, O. H.; Risselada, H. J.; Louhivuori, M.; Lindahl, E.; Vattulainen, I.; Marrink, S. J. 3D pressure field in lipid membranes and membrane-protein complexes. *Phys. Rev. Lett.* **2009**, *102* (7), 078101.

(5) Tepper, H. L.; Voth, G. A. A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *J. Chem. Phys.* **2005**, *122* (12), 124906.

(6) Knotts, T. A.; Rathore, N.; Schwartz, D. C.; de Pablo, J. J. A coarse grain model for DNA. *J. Chem. Phys.* **2007**, *126* (8), 084901.

(7) Chen, J.-S.; Teng, H.; Nakano, A. Wavelet-based multi-scale coarse graining approach for DNA molecules. *Finite Elem. Anal. Des.* **2007**, *43*, 346–360.

(8) Becker, N. B.; Everaers, R. From rigid base pairs to semi-flexible polymers: coarse-graining DNA. *Phys. Rev. E.: Stat. Nonlin. Soft. Matter Phys.* **2007**, *76* (2 Pt 1), 021923.

(9) Zhang, F.; Collins, M. A. Model simulations of DNA dynamics. *Phys. Rev. E* **1995**, *52* (4), 4217–4224.

(10) Mergell, B.; Ejtehadi, M. R.; Everaers, R. Modeling DNA structure, elasticity, and deformations at the base-pair level. *Phys. Rev. E* **2003**, *68*, 021911.

(11) Poulain, P.; Saladin, A.; Hartmann, B.; Prévost, C. Insights on Protein-DNA Recognition by Coarse Grain Modelling. *J. Comput. Chem.* **2008**, *29*, 2582–2592.

(12) Hyeon, C.; Thirumalai, D. Mechanical unfolding of RNA hairpins. *Proc. Natl. Acad. Sci. U. S. A* **2005**, *102* (19), 6789–6794.

(13) Hyeon, C.; Thirumalai, D. Forced-unfolding and force-quench refolding of RNA hairpins. *Biophys. J.* **2006**, *90* (10), 3410–3427.

(14) Zhang, D.; Konecny, R.; Baker, N. A.; McCammon, J. A. Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers* **2004**, *75* (4), 325–337.

(15) Forrey, C.; Muthukumar, M. Langevin Dynamics Simulations of Genome Packing in Bacteriophage. *Biophys. J.* **2006**, *91*, 25–41.

(16) Voltz, K.; Trylska, J.; Tozzini, V.; Kurkal-Siebert, V.; Langowski, J.; Smith, J. Coarse-grained force field for the nucleosome from self-consistent multiscaling. *J. Comput. Chem.* **2008**, *29* (9), 1429–1439.

(17) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure.* **2006**, *14* (11), 1633–1645.

(18) Tan, R. K. Z.; Petrov, A. S.; Harvey, S. C. YUP: A Molecular Simulation Program for Coarse-Grained and Multiscaled Models. *J. Chem. Theory Comput.* **2006**, *2*, 529–540.

(19) Korolev, N.; Lyubartsev, A. P.; Nordenskiold, L. Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails. *Biophys. J.* **2006**, *90* (12), 4305–4316.

(20) Wocjan, T.; Klenin, K.; Langowski, J. Brownian Dynamics Simulation of DNA Unrolling from the Nucleosome. *J. Phys. Chem. B* **2009**, *113* (9), 2639−2646.

(21) Langowski, J. Polymer chain models of DNA and chromatin. *Eur. Phys. J. E. Soft. Matter* **2006**, *19* (3), 241–249.

(22) Langowski, J.; Heermann, D. W. Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.* **2007**, *18* (5), 659–667.

(23) Arnott, S.; Campbell-Smith, P. J.; Chandrasekaran, R. Handbook of Biochemistry and Molecular Biology, 3rd Nucleic Acids ed.; CRC Press: Cleveland, OH, 1976; Vol. II, pp 411−422.

(24) 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.* **1996**, *100*, 19824–19839.

(25) *AMBER 10*; University of California: San Francisco, CA, 2008.

(26) 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* **1981**, *78* (4), 2179–2183.

(27) Dickerson, R. E.; Drew, H. R. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol.* **1981**, *149* (4), 761–786.

(28) Drew, H. R.; Dickerson, R. E. Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **1981**, *151* (3), 535–556.

(29) McConnell, K. J.; Beveridge, D. L. DNA structure: what's in charge. *J. Mol. Biol.* **2000**, *304* (5), 803–820.

(30) Phan, A. T.; Leroy, J. L.; Gueron, M. Determination of the residence time of water molecules hydrating B′-DNA and B-DNA, by one-dimensional zero-enhancement nuclear Overhauser effect spectroscopy. *J. Mol. Biol.* **1999**, *286* (2), 505–519.

(31) Young, M. A.; Ravishanker, D.; Beveridge, D. L. A 5-ns Molecular Dynamics Trajectory for B-DNA: Analysis of Structure, Motions, and Solvation. *Biophys. J.* **1997**, *73*, 2313–2336.

(32) Denisov, V. P.; Carlstrom, G.; Venu, K.; Halle, B. Kinetics of DNA hydration. *J. Mol. Biol.* **1997**, *268* (1), 118–136.

(33) Duan, Y.; Wilkosz, P.; Crowley, M.; Rosenberg, J. M. Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.* **1997**, *272* (4), 553–572.

(34) Pastor, R. W.; Brooks, B. R.; Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* **1988**, *65*, 1409–1419.

(35) Wu, X.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.

(36) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.

(37) Cheatham, T. E., III; Case, D. A. *Computational Studies of RNA and DNA*; Springer: Dordrecht, The Netherlands, 2006; pp 45−71.

(38) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(39) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., III; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92* (11), 3817–3829.

(40) Jorgensen, W. L. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

(41) Darden, T. A.; York, D.; Pedersen, L. Particle mesh Ewald: An $N\cdot\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(42) 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.* **1984**, *81*, 3684–3691.

(43) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(44) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6* (1), 63–91.

(45) 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.* **1981**, *78* (4), 2179–2183.

(46) Denisov, A. Y.; Zamaratski, E. V.; Maltseva, T. V.; Sandstrom, A.; Bekiroglu, S.; Altmann, K. H.; Egli, M.; Chattopadhyaya, J. The solution conformation of a carbocyclic analog of the Dickerson-Drew dodecamer: comparison with its own X-ray structure and that of the NMR structure of the native counterpart. *J. Biomol. Struct. Dyn.* **1998**, *16* (3), 547–568.

(47) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(48) Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry* **2004**, *43* (12), 3537–3554.

(49) Altan-Bonnet, G.; Libchaber, A.; Krichevsky, O. Bubble dynamics in double-stranded DNA. *Phys. Rev. Lett.* **2003**, *90* (13), 138101.

(50) Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., III; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **2010**, *38* (1), 299–313.

(51) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **2005**, *89* (6), 3721–3740.

(52) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **1999**, (101), 426–434.

(53) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(54) Piana, S. Atomistic simulation of the DNA helix-coil transition. *J. Phys. Chem. A* **2007**, *111* (49), 12349–12354.

(55) Cheatham, T. E., III; Kollman, P. A. Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **1996**, *259* (3), 434–444.

(56) Soliva, R.; Luque, F. J.; Alhambra, C.; Orozco, M. Role of sugar re-puckering in the transition of A and B forms of DNA

in solution. A molecular dynamics study. *J. Biomol. Struct. Dyn.* **1999**, *17* (1), 89–99.

(57) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(58) Gueron, M.; Leroy, J. L. Studies of base pair kinetics by NMR measurement of proton exchange. *Methods Enzymol.* **1995**, *261*, 383–413.

(59) Holbrook, S. R.; Kim, S. H. Local mobility of nucleic acids as determined from crystallographic data. I. RNA and B form DNA. *J. Mol. Biol.* **1984**, *173* (3), 361–388.

(60) Fujimoto, B. S.; Willie, S. T.; Reid, B. R.; Schurr, J. M. Position-dependent internal motions and effective correlation times for magnetization transfer in DNA. *J. Magn Reson. B* **1995**, *106* (1), 64–67.

(61) Kojima, C.; Ono, A.; Kainosho, M.; James, T. L. DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly 13C-labeled purine nucleotides. *J. Magn. Reson.* **1998**, *135* (2), 310–333.

(62) Andreatta, D.; Sen, S.; Pérez Lustres, J. L.; Kovalenko, A. E. N. P. M. C. J.; Coleman, R. S. B M. A. Ultrafast Dynamics in DNA: "Fraying" at the End of the Helix. *J. Am. Chem. Soc.* **2006**, *128*, 6885–6892.

(63) 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* (27), 7812–7824.

CT900653P

Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations

Ari Zeida, Matías Rodrigo Machado, Pablo Daniel Dans, and Sergio Pantano*
Institut Pasteur de Montevideo, Calle Mataojo 2020, Montevideo, Codigo Postal 11400, Uruguay
(Received 4 July 2011; revised manuscript received 1 June 2012; published 3 August 2012)

Bending of the seemingly stiff DNA double helix is a fundamental physical process for any living organism. Specialized proteins recognize DNA inducing and stabilizing sharp curvatures of the double helix. However, experimental evidence suggests a high protein-independent flexibility of DNA. On the basis of coarse-grained simulations, we propose that DNA experiences thermally induced kinks associated with the spontaneous formation of internal bubbles. Comparison of the protein-induced DNA curvature calculated from the Protein Data Bank with that sampled by our simulations suggests that thermally induced distortions can account for ∼80% of the DNA curvature present in experimentally solved structures.

DOI: 10.1103/PhysRevE.86.021903 PACS number(s): 87.14.gk, 87.15.A−, 87.15.H−

I. INTRODUCTION

The double-stranded DNA (dsDNA) polymer presents a very stable structure with persistence lengths on the order of 50 nm or ∼150 consecutive base pairs (bp). This stiffness results from the sum of relatively small energetic contributions (below $4k_BT$) [1], which come from the pairing and stacking of a large number of single nucleotides. Thermal excitation leads to the sporadic breaking of a single bp, giving place to the so-called "breathing" motion in which opening and closing can be detected by nuclear magnetic resonance (NMR) spectroscopy in the time range from picoseconds to nanoseconds [2]. The simultaneous opening of two or more base pairs is referred to as a "bubble". This kind of energetically more expensive event is, consequently, less frequent. Under room conditions, bubbles from 2 to 10 bp with characteristic closing times of several microseconds have been deduced from fluorescence quenching experiments [3]. This highly dynamic behavior, spanning about 8 orders of magnitude in time, challenges theoretical methods as molecular dynamics (MD) or wormlike polymer models to properly describe the flexibility of DNA. In this regard, coarse-grained (CG) models offer a valid alternative to achieve an exhaustive exploration of the conformational space of macromolecules with a significantly reduced computational effort, yet capturing the essential physics at play. Recent examples of CG models applied to the study of nucleic acids include the work of Savelyev and Papoian, who developed an accurate scheme using one effective bead per nucleotide. This model, which explicitly included ionic strength, allowed for the characterization of structural transitions and other DNA properties [4]; Ortiz and de Pablo [5] measured the effects of sequence on the stability of DNA under bending conditions via umbrella sampling calculations using a three beads per bp model [6]. At a higher level of abstraction, Alexandrov *et al.* [7] used a unidimensional nonlinear Langevin dynamics model to examine the breathing dynamics around the transcription start site of an engineered DNA segment, which works as a highly efficient transcriptional promoter called super core promoter 1 (SCP1) [8]. They inferred a relationship between intrinsic DNA flexibility and transcriptional activity.

Aimed at providing new insight on the flexibility of double-stranded DNA, we conducted a series of MD studies on DNA filaments at the CG level using the model reported by Dans *et al.* [9] (Fig. 1). Our scheme is somewhat reminiscent of the early model of Zhang and Collins [10]. Nucleotides are represented by six effective beads, each of them placed on the Cartesian coordinates of one element in the all-atom representation. These beads are characterized by particular masses, partial charges, bonding, and van der Waals parameters, so as to condense the molecular information from its atomic neighborhood in one effective interaction site. This translates into a reduction of nearly 80% in the number of particles constituting a given molecular system. In addition to predict melting temperatures under different conditions and hydrogen bonding features, our CG scheme accurately reproduces the structure and dynamics of dsDNA. Among many other accurate CG models of DNA, a distinctive feature of our CG scheme is the possibility to furnish fully atomistic structures from the CG trajectories thanks to its complete backmapping capabilities [9].

In this paper, we explore the unbiased dynamic behavior of single-stranded DNA (ssDNA) and double-stranded DNA in a multimicrosecond time scale through a series of simulations. Breathing motions, bubbles, kinks formation, and bending of DNA are studied for experimentally characterized molecular systems of two different sequences and lengths. Finally, we also compare our results with DNA curvature calculated from protein-dsDNA complexes in the Protein Data Bank (PDB).

II. METHODS

Our CG mapping scheme uses effective beads, which are placed on the positions of real atoms (Fig. 1). This offers the possibility to recover pseudoatomistic information from a CG trajectory by using internal coordinates defined for each atomistic nucleotide. Reconstruction of atomistic coordinates from CG configurations results in root mean square superposition on the order of 0.1 nm, which is well compared to the intrinsic variability observed during classical all-atoms simulations [9]. However, it is worth noting that our backmapping procedure tends to homogenize certain substates observed during the all-atom simulations. After backmapping, the ζ/ε torsions are always in the BI conformer, and α/γ torsions are reconstructed in the canonical $g-/g+$ distribution. Experimentally, nearly

*Corresponding author: spantano@pasteur.edu.uy

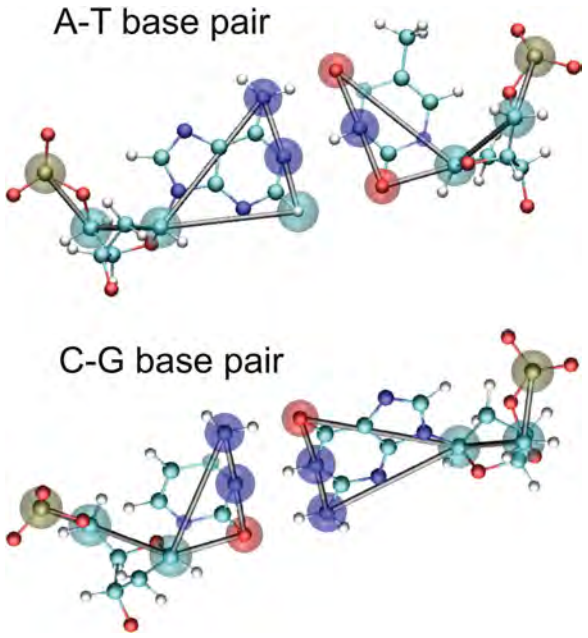

FIG. 1. (Color online) Coarse-grained model of DNA. Superposition of the CG and all-atom representations of the four nucleotides in a typical Watson-Crik interaction. Semitransparent spheres indicate the position of the atoms occupied by effective CG beads. Gray sticks indicate the connectivity between CG beads. See Refs. [9] and [11] for further details on the parameterization.

15% of the ζ/ε torsions are in the BII conformer, and some less frequent sequence-dependent shifts for the α/γ torsions are also observed. In our case, the sugar pucker is always reconstructed in the C2′-endo typical of the canonical B form. Nevertheless, as a result of the energy minimization performed as a final step of the backmapping procedure, we obtain 80% of C2′-endo conformations, whereas, the remaining conformers correspond mainly to C3′-exo.

To explore the dynamics of DNA, we choose different molecular systems. The first corresponds to the rationally designed SCP1, which sustains very high levels of transcription by RNA polymerase II [8]. The sequence of this dsDNA promoter is as follows:
5′-d(GCATGCCTGCAGGTACT**TATATAA**GGGGGTGGGG *GCGCGTTCGT*$_{(C)}$*C*$_{(G)}$CTCA$_{+1}$GTCGCGA$_{(G)}$ TCGAACA$_{(G)}$CTCGAGCCGAGCAGACGTGCCTACGGACCGTCTAG AGGATCC)-3′.

Bold letters indicate the AT-rich TATA box where the TATA binding protein binds. The underlined nucleotides correspond to the promoter region, whereas, the flanking sequences belong to the plasmid used in the experimental paper [8] and were included in the simulation to rule out possible end effects. The transcription starting site is indicated with a +1 subscript. Four mutations were introduced on the SCP1 sequence to study variations in the breathing profile as they reduce the transcription yield in *in vitro* experiments by 80% [7]. The subscripts between parentheses indicate the point mutations introduced in each of the precedent positions. The mutated SCP1 is named, hereafter, as mSCP1. MD simulations on systems SCP1 and mSCP1 were performed for 20 μs. Loose harmonic constraints of 3 kcal mol^{-1} Å^{-2} were used to mimic the continuation of the DNA within the plasmid.

Additionally, we also simulated a dsDNA segment analogous to the sequence called M18 in the paper by Altan-Bonnet *et al.* [3]. This corresponds to a dsDNA system with the following sequence:
5′-d(GGCGCCCAATATAAAATATTAAAATGCGC)-3′.

Constraints as those used for SCP1 and mSCP1 were used only at the 3′ end of the M18 dsDNA. This was intended to mimic the presence of a thymine tetraloop present in the experimental work, whose structure is unknown [3]. Five independent replicas of this system were simulated for 50 μs each using different starting conformers. This is equivalent to a total sampling of 250 μs.

Finally, two ssDNA, corresponding to the two separate filaments of M18, were simulated. Each of the two single-stranded filaments was simulated for 50 μs, corresponding to a total of 100 μs of sampling time for ssDNA.

All MD simulations were performed using AMBER 10 [12]. Temperature was controlled using a Langevin thermostat [13] with a collision frequency of 50 ps^{-1} and a target temperature of 300 K. Calculations were carried out under the same conditions reported by us in Ref. [9] with the parameter's modification reported in Ref. [11], which allows for a time step of 20 ps. Electrostatic interactions were calculated using a cutoff of 1.8 nm within the framework of the generalized Born model for implicit solvation as implemented in AMBER. Within this approach, electrostatic screening of a monovalent salt at a concentration of 150 mM was included via the Debye-Huckel parameter [14].

III. RESULTS AND DISCUSSIONS

Aimed at studying the intrinsic flexibility of DNA, we first analyzed the dynamics of the SCP1. This is a rationally designed transcriptional promoter of RNA polymerase II. It is a very well characterized system where the first protein binding event corresponds to the interaction between the TATA binding protein and its cognate DNA target motif. This eventually generates a cascade of protein DNA binding events, which ultimately results in high levels of transcriptional activity [8]. Theoretical methods have pointed out how breathing profiles are directly correlated with transcriptional regulation in this system [7]. MD simulations may further contribute to this by furnishing structural and dynamic insights. Moreover, the wide time window needed for a proper description of DNA dynamics makes CG simulations an attractive alternative.

The breathing profile of the SCP1 system, presented in Fig. 2(a), suggests a good conservation of the Watson-Crick hydrogen bond pattern. Despite the high prevalence of the canonical dsDNA conformation, two different kinds of events are present with different time scales: (i) Fleeting and widespread breathing profiles in the range of picoseconds to nanoseconds, (ii) long-lasting disruption of one or two consecutive bp with closing times on the order of 1 μs. The first and second kinds of events are represented by a punctuated orange (or gray in the printed version) and white patterns in Fig. 2(a), respectively.

Analysis of the trajectory indicates a large flexibility in the dsDNA filament [Fig. 2(b)], which is apparently uncorrelated

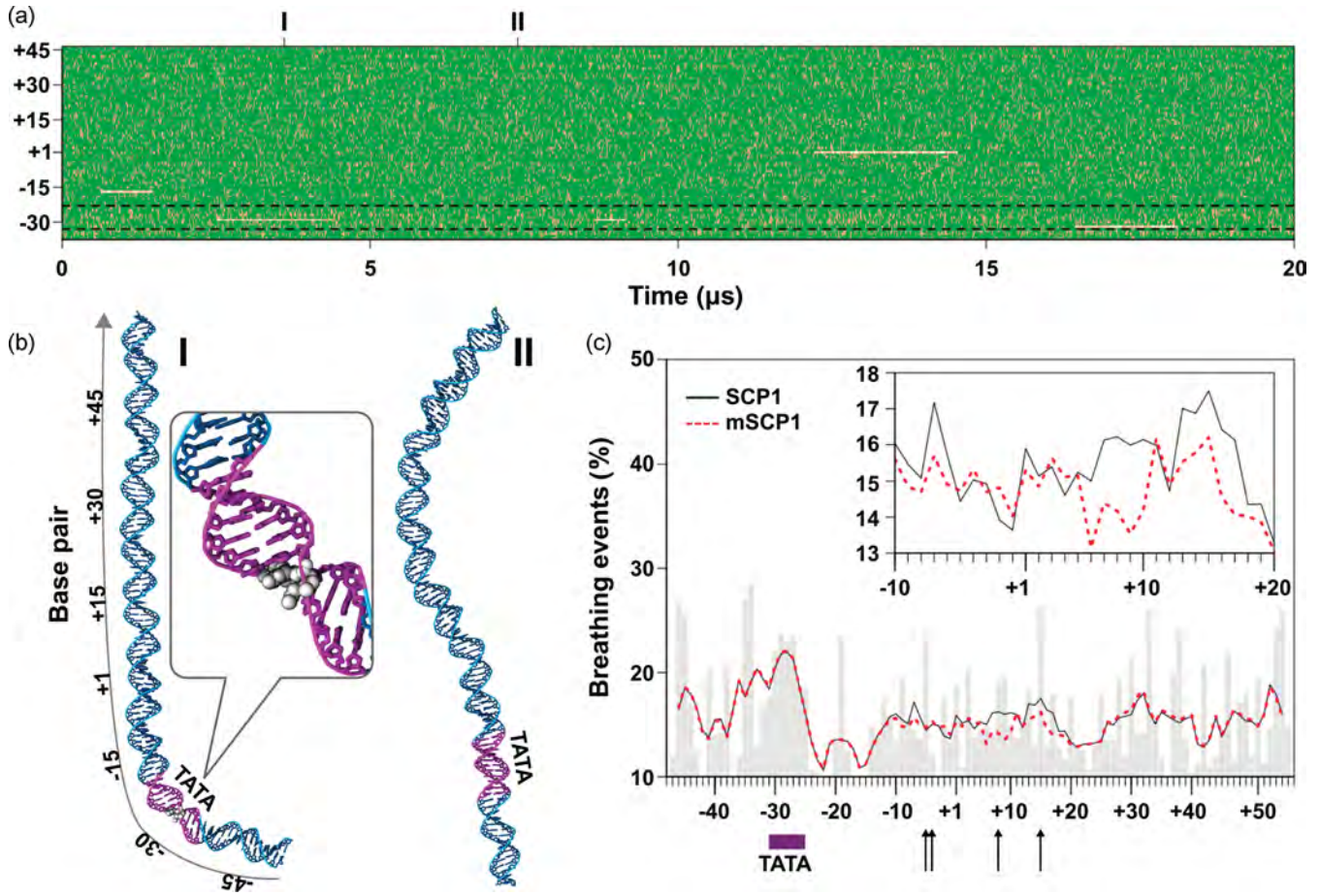

FIG. 2. (Color online) Simulations of SCP1 and mSCP1 dsDNA. (a) Breathing profile along the MD simulation of the SCP1 segment. Only the 80 bp corresponding to the promoter (underlined sequence in the Methods section) are shown for clarity. The distance between complementary bases in each bp is presented with different colors. Green (or dark gray in the printed version) color represents closed states (inter bp distances lower than 0.4 nm); open states are divided in two ranges: orange (light gray in the printed version): from 0.4 to 0.6 nm and white: more than 0.6 nm. The dashed lines delimit the AT-rich TATA box motif. (b) Representative conformers obtained along the trajectory to illustrate the flexibility of the filament. The roman numbers on top of panel (a) indicate the point in the trajectory where they come from. Inset: The single bp involved in the long-lasting disruption within the TATA box (conformer I) is shown with a space-filling representation. (c) Cumulative counting of the breathing events expressed as the percentage of time in which a bp is open between 0.4 and 0.6 nm. Gray bars are calculated for each bp in segments of the trajectory of the SCP1 filament. The continuous black line corresponds to an adjacent average every five data points. The dashed lines are calculated along the trajectory of mSCP1. The TATA box and the mutation sites are indicated by a rectangle and arrows, respectively. The inset shows a closeup on the region of the sequence where differences between SCP1 and mSCP1 are relevant.

with the fleeting breathing motion. However, we noticed that the disruption of even a single bp resulted in a marked curvature [Fig. 2(b)]. Therefore, we decided to characterize both events separately. With this aim, we calculated the cumulative sum of short breathing events of each individual bp excluding the segments of the trajectory where long-lasting disruptions are present [i.e., where the white color is present in Fig. 2(a)]. Although the fleeting breathing events may look apparently uncorrelated, the cumulative counting reveals that the AT-rich TATA box region experiences a more frequent breathing pattern. This seems to be a characteristic signature of the AT-rich tract since regions with higher nucleotide heterogeneity as the central segment, from positions -9 to $+15$, display a sensibly reduced breathing motion [Fig. 2(c)].

In contrast, the G-rich track immediately downstream of the TATA box, from positions -26 to -10, presents the lowest breathing occurrence. It is also worth noticing that the only peak within this region is centered on the thymine at position -19. This underlines the capacity of the model to pinpoint the effects of single nucleotides within a given sequence context.

It has been demonstrated that point mutations at positions -4, -5, $+8$, and $+15$, which are distant from the TATA box, are impaired severely in the transcription levels. Moreover, a correlation between transcriptional activity and a change in the opening probability around position $+1$ has also been inferred from a different simulation approach [7]. Therefore, we sought to perform a MD simulation of mSCP1. The quantification of the breathing pattern along the dynamics of mSCP1 showed differences with that of SCP1 only within regions separated, at most, 10 bp from the mutation site, i.e., in the neighborhood of the transcription starting site [Fig. 2(c)]. The breathing profile presented in Fig. 2(c), which can be related to the opening

probability of each bp, is in very good qualitative agreement with the description presented by Alexandrov *et al.* [7]. Our simulations are in line with their conclusion that the four point mutations modulate the dynamics around the transcription starting site, leaving unaffected different protein-DNA binding sites present in the promoter.

In addition to the breathing profiles, MD simulations also grant the possibility of acquiring structural insight. Along the dynamics, the SCP1 filament may experience a significant flexion [Fig. 2(b)]. Besides relatively short breathing movements on the order of the nanosecond, the simulation of the SCP1 also presented five long-lasting events scattered along the trajectory and the sequence. Three of these occur within the TATA box [Fig. 2(a)]. These events involve the opening of one or two consecutive bp with closing times from hundreds of nanoseconds up to nearly 3 μs. These more pronounced disruptions of the Watson-Crick pattern are related to a marked bending and, eventually, kinking [Fig. 2(b)]. Unfortunately, these long-lasting and apparently more relevant events happened in a time scale which is difficult to reach even for our CG scheme. Therefore, we sought to further explore this second kind of phenomenon in a smaller and computationally more affordable system, which allowed for longer simulation times. With this aim, we set up a series of simulations using the sequence M18 reported on in the Methods section. This sequence seems particularly well suited to our paper since it is relatively short and contains two GC "clamps" flanking an AT-rich track, which has been reported to favor the spontaneous formation of bubbles by fluorescence quenching experiments [3]. Moreover, significant breathing movements were previously reported by us for this system [9]. Hence, we speculated that a considerable increase in the sampling time could reveal a more complex behavior in terms of bubble formation and, eventually, more significant conformational changes. Therefore, we performed five independent simulations, each lasting 50 μs, i.e., a total sampling of 250 μs.

In analogy with the previous cases, the simulation of this shorter system revealed a complex behavior, which includes breathing, bubbling, and kinking. Moreover, we also observed the reversible separation of the 5′ end of the double helix (fraying), and partial melting-rehybridization [Fig. 3(a)]. In addition, the significantly longer sampling time helps to get a clearer discrimination of breathing and bubbling patterns, which are present at significantly different time scales [compare both panels of Fig. 3(a)]. Spontaneous and transient disruption of Watson-Crick interactions, involving 2–10 bp, occurred with no apparent correlation along the MD trajectories. The shortest temporal events are in the picosecond to nanosecond range in agreement with NMR measurements [2]. The overall integrity of the double helix is well preserved during these events [conformers I and II in Fig. 3(b)]. Additionally, a reduced number of bubbles spontaneously appear in the microsecond range [Fig. 3(a)].

Bubbles lasting picoseconds to nanoseconds do not translate into large conformational changes. However, during microsecond long bubbles, the separation between opposite phosphates was increased, on average, by ∼0.7 nm. The maximum increase in the interphosphate distance reached 1.7 nm but only for short periods (on the order of a few nanoseconds). This is consistent with the experimental determinations indicating that bubbles between 2 and 10 bp arise spontaneously at room temperature and under physiological salt conditions with lifetimes in the range of 20–100 μs [3]. It is uncertain, however, if these separations can fully account for the variations in the fluorescence quenching measured by Altan-Bonnet *et al.* [3].

The maximum temporal extension of the bubbles observed in our simulations is ∼3 μs in contrast with the nearly 1 order of magnitude higher bubbles deduced for this system from fluorescence quenching experiments [3]. This may suggest that a sampling time on the order of multimilliseconds may be needed to observe larger distortions in the backbone. In this context, it is important to recall the agreement between the lifetime of short-lived bubbles observed in our simulations and the NMR spectroscopy [2]. Although it is tempting to extrapolate this agreement to the range of the multimicroseconds, it is important to consider that the time scales sampled by CG simulations need to be, in general, interpreted with care.

A quantitative estimation of the hydrogen bond pattern (and, hence, on the stability of the double helix) can be acquired from the inter bp distances depicted in Fig. 4(a). Considering that a bp is formed at a distance below 0.4 nm [6,9], we observe that the first three bp are most frequently found in an open configuration. Furthermore, there is a rise in the inter bp distance near the middle of the double helix where long bubbles arise [compare also with Fig. 3(a)]. Besides the average and standard deviations, it is also worth paying attention to the extreme values. Although the minimum distances are limited by van der Waals contacts, relatively high maximum values are found, especially for bp numbers 18–21. Figure 4(a) provides a geometric view on the fact that, despite the global stability of the double helix, large distortions may spontaneously arise in the dsDNA at room conditions. These conclusions are in line with the results obtained by others using Langevin dynamics simulations [15].

To characterize the ordering in the open conformations, we calculated an order parameter, defined as $(\cos\ \langle\alpha_i\rangle)^2$, where α is the angle between the planes of individual bases between two consecutive nucleotides for all the i steps along the filament. This parameter provides an indication of the stacking between consecutive bases independent of the bp formation. The order parameter was calculated over the entire MD trajectories of M18 (i.e., 250 μs). In order to establish a comparison level, we also simulated two independent 50 μs long trajectories for ssDNA (see the Methods section). Moreover, to focus on the more distorted conformers within the double helix, we also selected 5 μs of the trajectory corresponding to the two bubbles centered on microseconds 13 and 22 on the trajectory presented in Fig. 3(a). The analysis was carried out independently for that selected piece of the trajectory since averaging over the entire trajectory flattens out the results.

The ssDNA filaments showed, as expected, a large variation corresponding to a highly unstructured molecule [Fig. 4(b)]. In stark contrast, the calculation on the dsDNA trajectories shows a low ordering only in the first two bp, which is indicative of the frequent but limited fraying at the extremity of the molecule. The order rapidly increases, converging to nearly unitary values already after the third base pair. The

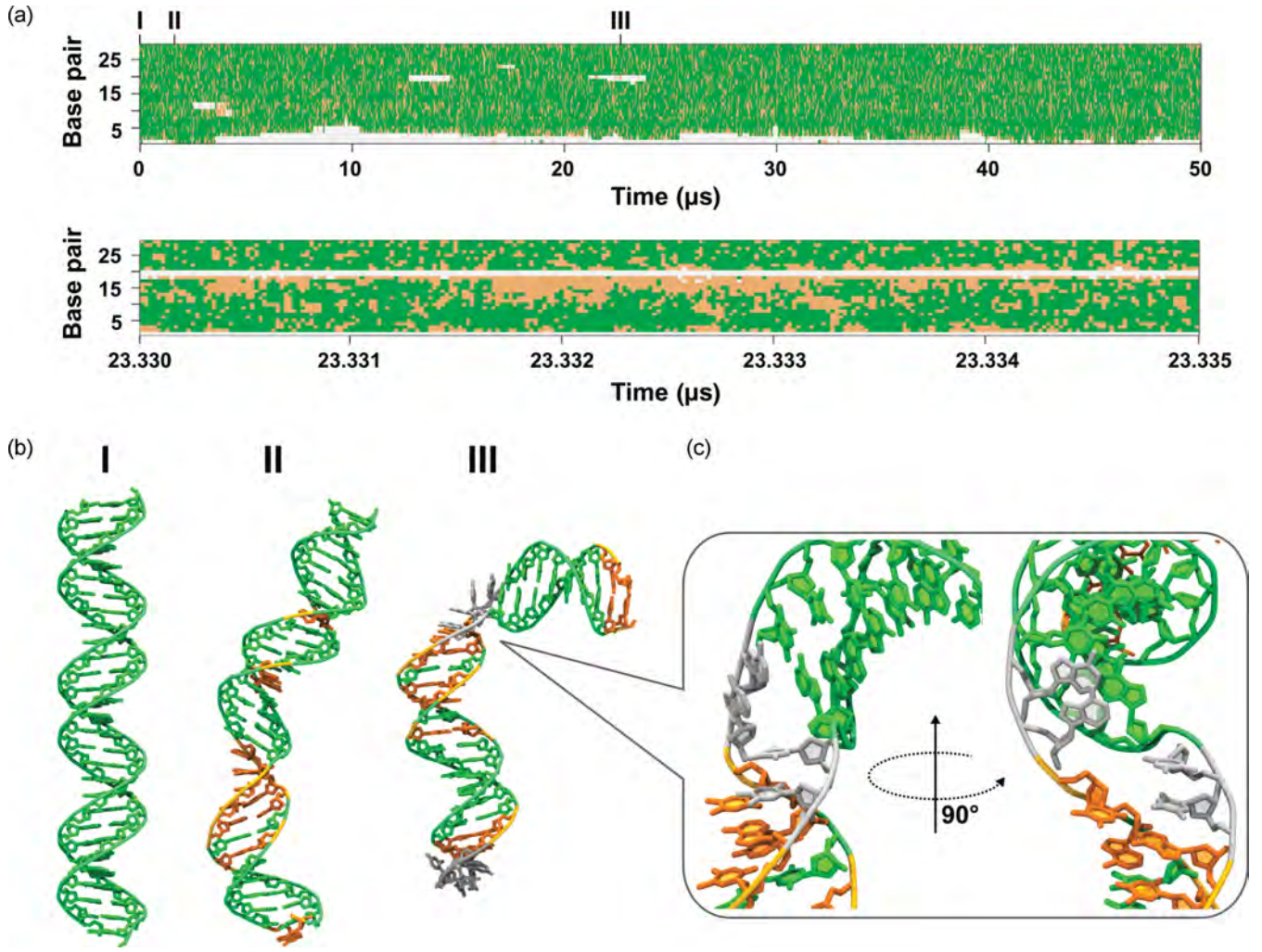

FIG. 3. (Color online) Simulations of M18 dsDNA. (a) Breathing profile along the MD simulation of the M18 segment. Color code corresponds to that of Fig. 2. Only one out of five independent 50 μs trajectories is shown for shortness. The bottom panel shows an inset of 5 ns within the 23rd microsecond where the fine structure of short-lived bubbles can be clearly appreciated. (b) All-atoms reconstituted (backmapped) molecular representations of three different representative conformers found during the simulations. The roman numbers indicate the point on the trajectory where they are taken from. Nucleotides are colored according to panel (a). (c) Two different orientations of a closeup on the kinked conformer shown in (b) III.

selected 5 μs trajectory containing bubbles showed a similar behavior with the exception of the highly bubbling region between bp 18 and 21. High levels of stacking between consecutive bases are kept even for bubbles with average base pair separations between 0.4 and 0.6 nm [Fig. 4(b)]. A significant decrease in the ordering is only observed when the base pair separation is, on average, higher than 0.6 nm. This supports the idea that locally denaturated regions retain some degree of order related to the base-base stacking [Fig. 3(c)], which would contribute to reduce the energy needed for bubble formation [3]. To exclude the possibility that the residual stacking in locally denaturated regions may arise from an artifact of our CG force field, we calculated the persistence length on the ssDNA filaments along the 100 μs of the trajectory. Since the structure of ssDNA is mainly stabilized by the stacking between contiguous bases and electrostatic repulsion between phosphate groups, parametrization defects would result in deviations in the persistence length from experimental results. The persistence length can be explicitly calculated from our simulations as $\Sigma^n \langle \cos \gamma_k \rangle$, where γ is the angle between the vectors perpendicular to the first and kth bases; the average is calculated over all the frames collected from the simulation, and the sum runs on all the n steps in the polymer. The persistence length of ssDNA resulted in 1.41 residues. This value is in very good agreement with experimental data [16], which suggest that the linearity of the ssDNA chain is completely lost already after only two residues.

Within the relatively large time window explored here, DNA breathing (but not bubbling) translates in a continuous flexion without compromising the integrity of the double helix. Clustering analysis of the curvature of DNA along the 250 μs of simulations showed that the most visited conformation is not a completely straight one. A histogram of the total bend of the dsDNA calculated using the program CURVES + [17] is shown in Fig. 4(c). We found that DNA bending follows

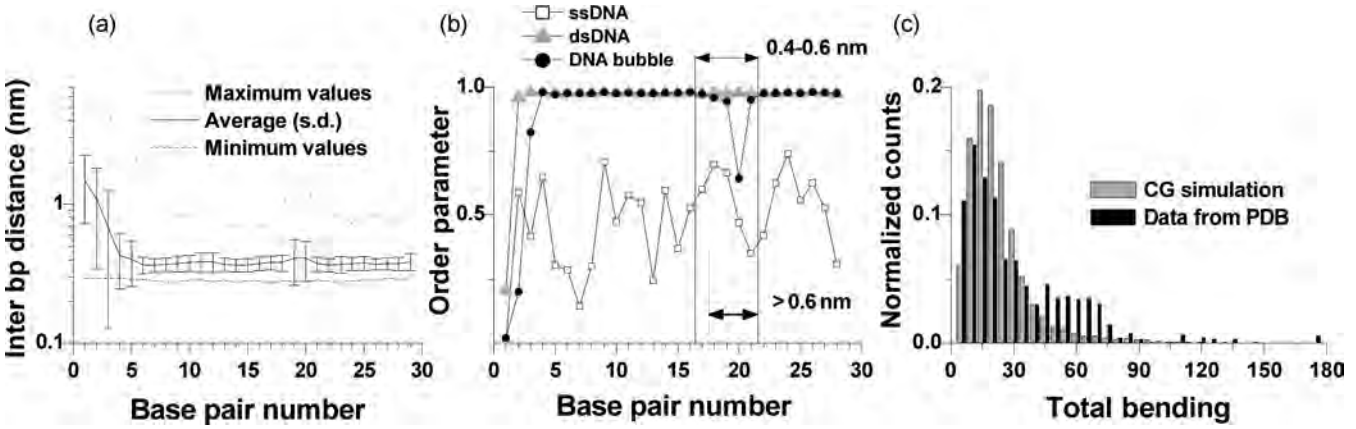

FIG. 4. Statistical analysis of DNA conformations on the M18 segment. (a) Continuous line: average inter bp distances as a function of the bp number. Standard deviations (s.d.) are reported as error bars. The maximum and minimum values measured during the dynamics are shown as dotted and dashed lines, respectively. For the sake of clearness, a logarithmic scale is used in the vertical axis. (b) Open squares: order parameter versus sequence calculated for ssDNA, gray triangles: dsDNA, and black circles: bubbled dsDNA trajectories. The arrows indicate the opening distance between bp in the region of the molecule where the bubbles occur. (c) Gray bars: histogram of the total bending of the dsDNA filament calculated over 250 μs of simulation. The black bars show the distribution of the total bend calculated for the protein-DNA complexes from the PDB (see main text).

an asymmetric distribution with a peak between 10° and 15°. Continuously curved DNA conformers can be found up to a maximum bend of ∼50°. Higher bending is only observed in the presence of bubbles, which may generate kinked conformations. Taking into account these kinked conformers, the bending distribution extends up to 130° [Fig. 4(c)]. This suggests that thermal fluctuations induce a number of conformations which can be roughly divided into three categories [Fig. 3(b)]: (i) nearly canonical and straight B-DNA, (ii) continuously bent DNA without significant bubbles in the double helix, and (iii) conformations with widely opened bubbles, which may generate kinked double helical filaments. The first two categories can be very well described by standard semiflexible polymer models [18], which predict a maximum thermally induced curvature of nearly 1 rad in double-stranded segments with extensions close to their persistence length. However, the kinked conformers arising in our simulations correspond to rare events with typical occurrence times on the microsecond scale.

The DNA deformability suggested by our results could be essential for defining its biophysical properties. In fact, sharp kinks related to spontaneous bubble formation have been proposed to explain DNA cyclization probabilities $>10^4$ times larger than those predicted by standard semiflexible polymer models for ∼100 bp long DNA segments [19] (i.e., shorter than their persistence length). Similar protein-independent DNA flexibility has also been reported using different experimental techniques [20–22]. According to a model proposed by Yan and Marko [23], thermal fluctuations are enough to generate hinges involving at least 3 bp long bubbles increasing the flexibility and the probability of cyclization. Contrasting interpretations for the high flexibility of DNA in terms of the disruption of Watson-Crick interactions or the continuous stacking between nucleobases has also been provided by Du *et al.* [24] and Geggier *et al.* [25]. The possibility of accessing nearly atomistic information from our CG simulations provides a structural picture for the idea that single bp openings may occur spontaneously in dsDNA. Bubbles may reach an extension up to ten consecutive bp with a disruption of the Watson-Crick hydrogen bond pattern, but residual ordering related with the base-base stacking is retained even in the bubbled region. When these distortions occur, they translate into a sharp kink in the double helix [Figs. 3(b) and 3(c)].

Simple geometrical calculations indicate that, at least, a single kink with an angle $\geqslant$120° plus the continuous bending of the rest of the chain is enough to join the two extremities of 100 bp dsDNA filaments in a teardrop shape. The probability of the occurrence of kinks higher than 120° calculated from the distribution of the bending angles [Fig. 4(c)] is on the order of 10^{-4}, supporting the spontaneous occurrence of rare conformational defects in DNA as responsible for cyclization of relatively short DNA segments.

The biological relevance of this phenomenon is highlighted if we compare our results with the bending measured from protein-dsDNA complexes reported in the Protein Data Bank. With this aim, we considered all the protein-dsDNA complexes solved at a resolution higher than 0.25 nm as taken from the human curated protein DNA interface database (http://melolab.org/pdidb) [26]. There is a fairly good correspondence between the angular distributions of free DNA (from our simulations) and the set of nearly all the experimentally determined protein-dsDNA complexes [Fig. 4(c)].

It also has to be taken into account that sequence-dependent effects are expected to modulate DNA deformability [27]. Recently, using another CG model combined with umbrella sampling techniques, Ortiz and de Pablo showed how sequence can change the ability of DNA to form kinks in nucleosome positioning segments [5]. Although sequence-dependent effects have not been explored in this paper (neither in the simulations nor in the analysis of the experimental DNA-protein complexes), our results suggest that thermally driven bending and metastable kinking of DNA is an intrinsic property of the double helical architecture.

IV. CONCLUSIONS

To summarize, the set of simulations presented here provide nearly atomistic details of the dynamics of DNA in the time scale of the multimicroseconds, yet unexplored. The simulation of the SCP1 system presents breathing profiles, which are consistent with transcription experiments and other theoretical methods. In addition, we obtain the indication that even one single bp disruption may originate marked kinks in the double helical filament. Comparison with mSCP1 suggests that the effect of point mutations can propagate up to 10 pb in the double helix.

The relatively long simulations of the M18 sequence highlight the relevance of long bubbles spontaneously appearing in dsDNA. These bubbles of up to 10 bp may originate marked kinks in the double helix. These kinks present noncanonical but still ordered structural motifs, which are stable in the multimicrosecond time scale. The relatively long lifetime of these conformers could have a deep biological relevance. In fact, comparison of the bending produced by spontaneously arising kinks with most of the experimentally determined protein-dsDNA complexes shows good correspondence. This suggests that thermally induced deformation of the double helix could be sufficient to overcome the free energy barrier needed to obtain about 80% of the DNA-protein complexes currently known.

ACKNOWLEDGMENTS

We thank J. J. Cifuentes and F. Melo for help with the analysis of the information from Protein-DNA interactions Database. 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 No. 8 DCI-ALA/2007/19.040 between Uruguay and the European Commission). M.R.M. is supported by a fellowship from CSIC-UdelaR. P.D.D. and S.P. appreciate support from the National Scientific Program of ANII (SNI) and from the Basic Science Development Program of Uruguay (PEDECIBA).

[1] J. SantaLucia, Jr., Proc. Natl. Acad. Sci. USA **95**, 1460 (1998).
[2] M. Gueron and J. L. Leroy, Methods Enzymol. **261**, 383 (1995).
[3] G. Altan-Bonnet, A. Libchaber, and O. Krichevsky, Phys. Rev. Lett. **90**, 138101 (2003).
[4] A. Savelyev and G. A. Papoian, Proc. Natl. Acad. Sci. USA **107**, 20340 (2010).
[5] V. Ortiz and J. J. de Pablo, Phys. Rev. Lett. **106**, 238107 (2011).
[6] T. A. Knotts, N. Rathore, D. C. Schwartz, and J. J. de Pablo, J. Chem. Phys. **126**, 084901 (2007).
[7] B. S. Alexandrov, V. Gelev, S. W. Yoo, L. B. Alexandrov, Y. Fukuyo, A. R. Bishop, K. O. Rasmussen, and A. Usheva, Nucleic Acids Res. **38**, 1790 (2010).
[8] T. Juven-Gershon, S. Cheng, and J. T. Kadonaga, Nat. Methods **3**, 917 (2006).
[9] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, J. Chem. Theory Comput. **6**, 1711 (2010).
[10] F. Zhang and M. A. Collins, Phys. Rev. E **52**, 4217 (1995).
[11] L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, and S. Pantano, J. Chem. Theory Comput. **6**, 3793 (2010).
[12] D. A. Case *et al.*, AMBER 10, University of California, San Francisco, 2008.
[13] R. W. Pastor, B. R. Brooks, and A. Szabo, Mol. Phys. **65**, 1409 (1988).
[14] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case, Theor. Chem. Acc. **101**, 426 (1999).
[15] O. C. Lee, J. H. Jeon, and W. Sung, Phys. Rev. E **81**, 021906 (2010).
[16] S. B. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).
[17] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, Nucleic Acids Res. **37**, 5917 (2009).
[18] P. J. Hagerman, Annu. Rev. Biophys. Biophys. Chem. **17**, 265 (1988).
[19] T. E. Cloutier and J. Widom, Mol. Cell **14**, 355 (2004).
[20] N. A. Becker, J. D. Kahn, and L. J. Maher III, J. Mol. Biol. **349**, 716 (2005).
[21] C. Yuan, H. Chen, X. W. Lou, and L. A. Archer, Phys. Rev. Lett. **100**, 018102 (2008).
[22] P. A. Wiggins, T. van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker, and P. C. Nelson, Nat. Nanotechnol. **1**, 137 (2006).
[23] J. Yan and J. F. Marko, Phys. Rev. Lett. **93**, 108108 (2004).
[24] Q. Du, A. Kotlyar, and A. Vologodskii, Nucleic Acids Res. **36**, 1120 (2008).
[25] S. Geggier, A. Kotlyar, and A. Vologodskii, Nucleic Acids Res. **39**, 1419 (2011).
[26] T. Norambuena and F. Melo, BMC Bioinf. **11**, 262 (2010).
[27] A. Perez, F. J. Luque, and M. Orozco, Acc. Chem. Res. **45**, 196 (2012).

JCTC Journal of Chemical Theory and Computation

Another Coarse Grain Model for Aqueous Solvation: WAT FOUR?

Leonardo Darré,[†] Matías R. Machado,[†] Pablo D. Dans,[†] Fernando E. Herrera,[†,‡] and Sergio Pantano*,[†]

Institut Pasteur de Montevideo, Calle Mataojo 2020, CP 11400, Montevideo, Uruguay, and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Avda. Rivadavia 1917 - CP C1033AAJ - Cdad. de Buenos Aires, Argentina

Received July 7, 2010

Abstract: Biological processes occur on space and time scales that are often unreachable for fully atomistic simulations. Therefore, simplified or coarse grain (CG) models for the theoretical study of these systems are frequently used. In this context, the accurate description of solvation properties remains an important and challenging field. In the present work, we report a new CG model based on the transient tetrahedral structures observed in pure water. Our representation lumps approximately 11 WATer molecules into FOUR tetrahedrally interconnected beads, hence the name WAT FOUR (WT4). Each bead carries a partial charge allowing the model to explicitly consider long-range electrostatics, generating its own dielectric permittivity and obviating the shortcomings of a uniform dielectric constant. We obtained a good representation of the aqueous environment for most biologically relevant temperature conditions in the range from 278 to 328 K. The model is applied to solvate simple CG electrolytes developed in this work (Na^+, K^+, and Cl^-) and a recently published simplified representation of nucleic acids. In both cases, we obtained a good resemblance of experimental data and atomistic simulations. In particular, the solvation structure around DNA, partial charge neutralization by counterions, preference for sodium over potassium, and ion mediated minor groove narrowing as reported from X-ray crystallography are well reproduced by the present scheme. The set of parameters presented here opens the possibility of reaching the multimicroseconds time scale, including explicit solvation, ionic specificity, and long-range electrostatics, keeping nearly atomistic resolution with significantly reduced computational cost.

Introduction

Computer simulation of biological systems is continuously experiencing a tremendous expansion urged by the ever-growing computer power that allows for the treatment of always more complex systems and for time scales that continuously approach biological relevancy.[1] Parallel to this, the greediness to achieve structural and dynamical descriptions of yet longer and bigger sized systems has prompted the scientific community to develop simplified models of molecular assemblies that mimic arbitrarily intricate molecular systems with a lower degree of complexity. These simplified or coarse grain (CG) representations reduce significantly the computational demands but still capture the physical essence of the phenomena under examination.[2,3] Starting from the pioneering simplified models used to describe protein folding,[4,5] a huge number of successful applications covering a wide range of biomolecular and nanotechnologically relevant applications have been presented.[6−18] For an exhaustive review of this area, the book *Coarse-Graining of Condensed Phase and Biomolecular Systems*[19] is recommended. In this context, the accurate treatment of solvent effects is still a challenging issue. In fact, many CG approaches use a uniform dielectric constant, which may produce an incorrect partition of

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

[†] Institut Pasteur de Montevideo.

[‡] CONICET.

10.1021/ct100379f
Published on Web 11/04/2010

hydrophilic molecules in a hydrophobic medium. Recently, elaborated and/or systematic developments of CG models for simulating water, Hbond (hydrogen bond) bound, and/or ionic liquids with high accuracy have been presented.[20−23] Here, we present a new and simple CG model for water derived from elementary physicochemical concepts and fitting the interaction parameters to reproduce some characteristic features of liquid water. The main advantage of our model is that all of the interactions are described by a typical Hamiltonian for classical simulations, explicitly including long-range electrostatics. This model is composed of four interconnected beads arranged in a tetrahedral conformation (Figure 1). Each bead carries an explicit partial charge. In this way, the liquid generates its own dielectric permittivity, avoiding the use of a constant dielectric medium. The model achieves a reasonable reproduction of some common properties of liquid water in the range of temperatures relevant for most biological applications.

As examples of the potentiality of the model, we study first the solvation of CG monovalent electrolytes developed in this work (Na^+, K^+, and Cl^-). Then, we present molecular dynamics (MD) simulations of a recently published CG model for DNA.[24] This model was shown to provide nearly atomistic resolution information of the structure and dynamics of double-stranded DNA under the generalized Born model approach for implicit solvation. In this contribution, we present an extension of that model for explicit solvation.

We show that this CG scheme is able to reproduce solvation spines, electrolyte specificity, and cation-driven narrowing of the minor groove. These examples illustrate the usefulness of the model in incorporating electrostatic effects in a physiological medium, keeping the chemical details of the different ionic species within CG simulations and overcoming the drawbacks of implicit solvation.

Methods

Description of the Model. The underlying idea of the model is that, due to its molecular characteristics, pure water behaves as a structured liquid forming (among other structural arrangements) transient tetrahedral clusters.[25] These clusters are composed of a central water coordinated by four identical molecules that form an elementary tetrahedral arrangement (Figure 1A). In this arrangement, the central molecule is buried and unable to interact with any other particle outside of the cluster. Our working hypothesis is that, owing to the replication of this structure in the bulk, the central molecule of any tetrahedron can be taken into account implicitly passing from a highly packed (atomistic) to a more granular (CG) liquid (Figure 1B). Aimed at reproducing the structural organization of the liquid, we generated a molecular topology in which four "covalently bound" beads are placed on the geometric positions of four oxygen atoms at the corners of an ideal tetrahedron (hence, the name WAT-FOUR or WT4 for short). The proposed topology implies that within an elementary cluster, the Hbond interactions that hold together the atomistic liquid water are represented by spring constants linking four beads (Figure 1B). The interactions between elementary clusters are taken into account by normal vdW and electrostatic terms in the classical Hamiltonian (Table S1, Supporting Information). These forces reproduce the overall tetrahedral ordering of water, allowing the elementary clusters to diffuse freely.

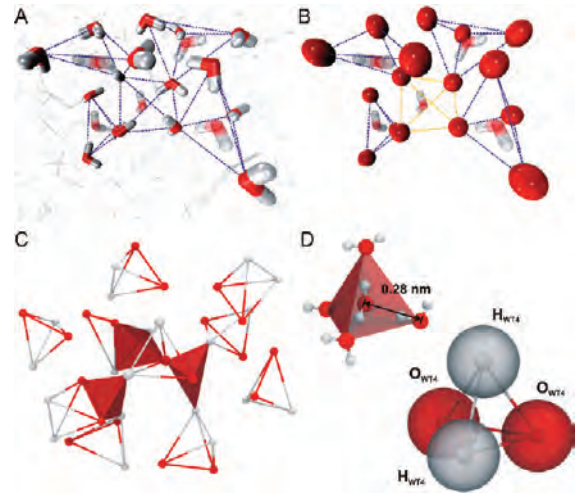

Figure 1. From atomistic to CG water. (A) Snapshot taken from a MD simulation showing the typical ordering of bulk water molecules. Gray molecules represent the liquid bulk. The structural organization is illustrated with a few opaque, thick water molecules which occupy the corners of irregular tetrahedrons. They saturate the Hbond capacity of a (semitransparent) molecule located in the center of each tetrahedron. Hbonds are indicated with dashed lines. (B) The positions of each of the oxygen atoms at the corners of the tetrahedra in A are now indicated with red beads. The concept behind the WAT FOUR (WT4) model is that those elementary tetrahedral clusters can be represented by four harmonically linked beads. The covalent bonds included in the WT4 model are represented by dark dashed lines, while intercluster interactions (vdW and electrostatics) are indicated with light dashed lines. The model implies that a number of water molecules are taken into account implicitly (represented as semitransparent molecules). Notice that water molecules can be implicitly represented even between noncovalently bound beads (take, for example, the central water molecule in the picture). The positions of all of the elements in A and B are identical. (C) Structural organization of WT4 in the bulk solution taken from a MD snapshot. The model reproduces higher-granularity tetrahedral organization in the space through noncovalent interactions. Red planes evidence the presence of rough tetrahedrons formed between different WT4 molecules comprising an implicit water molecule. (D) The ideal organization of a tetrahedral water cluster leads to the geometry of WT4. The separation of 0.28 nm between the water oxygen located at the center of the tetrahedron and its corners corresponds to the oxygen−oxygen (first neighbor) distance. This elementary cluster can be mapped to a WT4 molecule (bottom) composed by four harmonically bonded beads. White and red beads (hydrogen-like, H_{WT4}, and oxygen-like, O_{WT4}) carry positive and negative partial charges of 0.41*e*, respectively.

In analogy with the nearly tetrahedral water molecule that promotes a tetrahedral ordering in the surrounding space, a WT4 molecule recreates a roughly similar arrangement with a higher granularity (Figure 1C). Indeed, the structure of a WT4 molecule is replicated in its neighborhood, leaving holes that can be regarded as atomistic waters implicitly taken

Table 1. Interaction Parameters of the CG Models for Water and Ions[a]

					bond parameters	
	mass (au)	charge (*e*)	σ^b (nm)	ε (kJ mol^{-1})	d_{eq} (nm)	$K_{bond}{}^c$ (kJ mol^{-1} nm^{-2})
SPC[27]	Ow:16	Ow:−0.82	0.3166	0.650	0.1[d]	172500
	Hw:1	Hw:+0.41				
TIP3P[28]	Ow:16	Ow:−0.8340	0.315061	0.6364	0.09572[d]	251208
	Hw:1	Hw:+0.4170				
WT4	O_{WT4}:50	O_{WT4}:−0.41	0.42	0.55	0.45[e]	2092
	H_{WT4}:50	H_{WT4}:+0.41				
NaW^+	130.99	1	0.58	0.55		
KW^+	147.1	1	0.645	0.55		
ClW^-	143.45	−1	0.68	0.55		

[a] The parameters of two common atomistic water models (SPC and TIP3P) are included for comparison. [b] Distance from the atomic center to the minimum of the vdW function. [c] Corresponds to a harmonic approximation of the form $E_{bond} = K_{bond}(d - d_{eq})^2$. [d] Hydrogen−oxygen distance. [e] Interbead distance.

into account by the CG scheme. These implicit waters can be present not only within the four bonded beads but also between tetrahedrons formed by beads belonging to a different molecule (Figure 1B and C). This suggests that the WT4 molecules in the bulk solution have the capacity to form interactions alike to Hbond networks.

The distance between the central oxygen of a tetrahedral water cluster (Figure 1D) and any other oxygen is ∼0.28 nm, as determined from diffraction experiments.[26] Taking this into account and the geometry of a perfect tetrahedron, the equilibrium distance between beads was set to 0.45 nm. The bond stretching force constant was set to mimic the interaction energy involved in typical hydrogen bonds. We tried harmonic constants within a range from 837 kJ/mol nm^2 to 4184 kJ/mol nm^2 (2 kcal/mol Å^2 to 10 kcal/mol Å^2). A value of 2092 kJ/mol nm^2 (5 kcal/mol Å^2) was chosen, as it results in a better fit of different water properties. This weak link confers the molecule a certain degree of structural plasticity, resulting in small deviations from a perfect tetrahedron upon temperature effects. These deformations could be identified with the nonperfect tetragonal ordering present in liquid water at room temperature. Given the tetrahedral symmetry, only these two bonded parameters for intramolecular interactions are needed (Table 1 and Figure 1D).

Intermolecular nonbonded interactions are ruled by normal van der Waals and electrostatic parameters, listed in Table 1. Partial charges were assigned considering that the central water molecule in a given atomistic tetrahedral cluster neutralizes the atomic charges of the waters in the corners by Hbond formation. If the water atomic charges are q for the hydrogen and $-2q$ for the oxygen, this yields two positive corners with charge q (alike to Hbond acceptors) and two negative corners with charge $-q$ (alike to Hbond donors, Figure 1D). The assignment of partial charges is a largely unsolved issue in classical force fields. In the particular case of water, this task has been addressed in many different ways, from adjusting parameters to reproduce experimental quantities in the liquid or gas phase to ab initio potentials derived from calculations using small clusters of molecules. However, no available model is capable of reproducing all of the water properties with good accuracy. Given the roughness of our model, we just sought to keep the electrostatic interactions engaged by CG beads comparable to atomistic Hbonds. Therefore, we simply tried the same atomic charge values used in common three-point water models (Table 1). Among several atomistic three-point water models tried, the charge distribution that better fit the experimental values was that of the SPC model.[27] The van der Waals radii and well deepness were used as free parameters. Intramolecular nonbonded interactions were excluded.

The mass of each bead was assigned to fit the density of liquid water. To this task, we used a computational box containing 497 WT4 molecules simulated at 300 K and 1 bar. The mass per bead necessary to match a density close to 1 g/mL resulted in 50 au. Taking into account that the mass of each atomistic water molecule is 18 au, it is implied that each WT4 bead represents ∼2.8 water molecules (50 au/18 au). This corresponds on average to about 11 real waters per WT4 molecule. Namely, we assume that each WT4 molecule represents 11 real water molecules in the CG scheme. Therefore, whenever we compare with physicochemical properties, a renormalization factor of 11 is taken into account (see below).

The packing factor of the WT4 spheres calculated as the volume of the cubic box that contains the WT4 molecules divided into the excluded volume of beads is ∼0.47, close to the 0.5 calculated for the SPC model. These values are significantly lower than the ideal 0.74 expected for the hexagonal closest packing (the maximum compaction for rigid spheres). This suggests that the bulk structure of WT4 leaves a number of interstitial cavities in a slightly higher proportion than in the SPC model.

CG Model for the Ions. Three ionic species were developed to represent, at the CG level, the hydrated states of Na^+, K^+, and Cl^- (hereafter called NaW^+, KW^+ and ClW^-, respectively).

Ions were developed considering that six water molecules are always attached to them[29] (i.e., roughly considering an implicit first solvation shell). Therefore, their masses were set as the sum of the ionic mass plus that of six water molecules (Table 1). Partial charges were set to unitary values. The van der Waals radii were adopted to match the first minima of the radial distribution function (RDF, also known as $g(r)$) of hydrated ions as obtained from neutron diffraction experiments.[30] The deepness of the well was set to the same values as the WT4 beads. This was done to ensure compatibility since when a WT4 molecule contacts a CG ion it interacts with its first solvation shell,

Table 2. Description of the Simulated Systems

system		water model	number of molecules	ionic species (number of ions)[a]	solute	temperature (K)	simulation time (ns)	ionic pair concentration (M)
AA[b]	S_1^{AA}	SPC	2483[c]			278−323	45	
AA	S_2^{AA}	SPC	5368[c]	Na^+(1) Cl^-(1)		300	20	0.01
AA	S_3^{AA}	SPC	5368[c]	K^+(1) Cl^-(1)		300	20	0.01
AA	S_4^{AA}	TIP3P	2483[c]			278−323	45	
AA	S_5^{AA}	TIP3P	7612[c]	Na^+(34) Cl^-(34)		300	15	0.5
CG[d]	S_1^{CG}	WT4	497[e]			300	100	
CG	S_2^{CG}	WT4	497[e]			278−328	200	
CG	S_3^{CG}	WT4	268[e]			300	3	
CG	S_4^{CG}	WT4	268[e]			300	3	
CG	S_5^{CG}	WT4	473[e]	NaW^+(1) ClW^-(1)		300	100	0.01
CG	S_6^{CG}	WT4	473[e]	KW^+(1) ClW^-(1)		300	100	0.01
CG	S_7^{CG}	WT4	456[e]	NaW^+(44) ClW^-(44)		300	30	0.5
CG	S_8^{CG}	WT4	456[e]	KW^+(44) ClW^-(44)		300	30	0.5
CG	S_9^{CG}	WT4	174[e]	NaW^+(7) ClW^-(7)		300	200	0.2
CG	S_{10}^{CG}	WT4	174[e]	KW^+(7) ClW^-(7)		300	200	0.2
CG	S_{11}^{CG}	WT4	170[e]	NaW^+(11) ClW^-(11)		300	200	0.3
CG	S_{12}^{CG}	WT4	170[e]	KW^+(11) ClW^-(11)		300	200	0.3
CG	S_{13}^{CG}	WT4	655[e]	NaW^+(34) ClW^-(34)		300	100	0.5
CG	S_{14}^{CG}	WT4	655[e]	KW^+(34) ClW^-(34)		300	100	0.5
CG	S_{15}^{CG}	WT4	506[e]	NaW^+(19) KW^+(19) ClW^-(16)	CG-DNA	300	4000	0.15[f]

[a] Parameters from Berendsen et al.[44] and van Gunsteren et al.[45] In system S_5^{AA}, the CHARMM PARAM27 parameters[46] were used. [b] AA: all atoms. [c] Atomistic water molecules. [d] CG: coarse grain. [e] WT4 molecules. [f] Not considering 22 neutralizing counterions.

which is implicitly considered. A list of nonbonded interaction parameters for the CG monovalent ions is detailed in Table 1.

CG Model for DNA. The CG system used for DNA was essentially the same as that previously presented by us.[24] This CG model reduces the complexity of the atomistic picture to six beads per nucleobase (see Supporting Information Figure S1 for the coarse graining scheme). This mapping keeps the "chemical sense" of specific Watson−Crick recognition allowing the 5′−3′ polarity. Similarly to the approach taken here for water and ions, molecular interactions are evaluated using a classical Hamiltonian. The beads used in this representation carry partial charges, which permits the use of explicit electrostatics

Minor changes have been introduced to the interaction parameters to improve the stability of the double strand using a time step of 20 fs. Back mapping of the atomic coordinates during the trajectory permitted an evaluation of the overall structural quality of the DNA dodecamer in terms of helical parameters (Supporting Information Figure S2). This new parameter set reproduces equally well the structural features of the double-stranded helix.

The complete set of new parameters for DNA is listed in Supporting Information Table S1.

A similarity index between the present implementation and that using the GB model for implicit solvation was calculated from the covariance matrices obtained from the trajectories performed in the present work and that performed in Dans et al.[24] for the Drew−Dickerson dodecamer.

Molecular Dynamics. MD simulations were performed using Gromacs 4.0.5[31−34] in the NPT ensemble unless otherwise stated. The temperature was coupled using the Nose−Hoover thermostat,[35,36] while pressure was kept at 1 bar by means of a Parrinello−Rahman[37,38] barostat, with coupling times of 1 and 5 ps, respectively. A cutoff for nonbonded interactions of 1.2 nm was used, while long-range electrostatics were evaluated using the Particle Mesh Ewald approach.[39,40] A time step of 2 fs was used in all-atom (AA) simulations, while in the CG simulations the time step was set to 20 fs. In order to ensure that the use of such a relatively long integration step does not introduce energy conservation problems, we performed a series of simulations at constant energy (NVE ensemble) using such a time step and varying the cutoff. For an acceptable accuracy in the integration of the equations of motion, one should expect the fluctuations of the total energy to be lower than one-fifth (20%) of the kinetic or potential energy components of the system.[41] According to our results, this criterion is well fulfilled with total energy fluctuations representing 5% of potential or kinetic energy fluctuations, using cutoff values of 1.0, 1.2, and 1.5 nm (Supporting Information Table S2). It was decided to use a cutoff of 1.2 nm, which besides ensuring energy conservation also includes direct nonbonded interactions up to the second neighboring WT4 molecule in solution. Additionally, NVT simulations were performed for some systems in order to compute the WT4 surface tension and the ionic osmotic pressure as detailed below.

All of the interactions (i.e., WT4−WT4, WT4−ion, ion−ion, ion−DNA, WT4−DNA, and DNA−DNA) were straightforwardly calculated within the pairwise Hamiltonian of Gromacs 4.0.5, which is common to many popular MD packages. The van der Waals cross interactions were calculated using the Lorentz−Berthelot combination rules.

Five atomistic (S_{1-5}^{AA}) and 15 CG systems (S_{1-15}^{CG}) were constructed to evaluate different properties of interest (see Table 2). Atomistic simulations were used to obtain reference properties to be compared with the CG models for water and ions. Systems S_1^{AA} and S_4^{AA} were used to compute density and diffusion coefficient profiles in a relevant range of temperatures (see Table 2). The temperature scan was carried out raising the reference temperature by 5° in steps of 5 ns.

Both radial distribution functions (ion−Ow) and electrostatic potential (on the line connecting both ions) were

calculated from systems S_2^{AA} and S_3^{AA}, where the cation−anion distance was kept fixed at 3.6 nm during the whole simulation. This last property was also calculated for system S_1^{AA} at room temperature in order to use it as a reference state for pure water. System S_5^{AA} was used to validate the methodology for measuring the osmotic pressure (described in the Supporting Information).

Regarding the CG simulations, bulk water properties under room conditions were obtained from system S_1^{CG}. The behavior of the model in the range of temperatures from 278 to 328 K was assessed using system S_2^{CG}. The temperature scan was carried out as in the corresponding atomistic simulations (S_1^{AA} and S_4^{AA}) but using time windows of 20 ns instead of 5 ns.

Surface tension and isothermal compressibility at the CG level were computed from systems S_3^{CG} and S_4^{CG}, respectively, according to the following steps, as proposed elsewhere.[42] First, an initial configuration at 300 K and 1 bar (generated by a short NPT equilibration of a simulation box containing 268 WT4 molecules) underwent a 0.1 ns equilibration in the NVT ensemble. The resulting configuration was used, on one hand, to construct system S_3^{CG} by adding vacuum slabs above and below the water bulk, so the box length in the *z* direction was tripled. A 3 ns production NVT simulation was conducted in such a system at 300 K, from which the surface tension was computed from the pressure tensors:

$$\gamma = \frac{L_z}{2}\left\langle P_{zz} - \left(\frac{P_{xx} + P_{yy}}{2}\right)\right\rangle \quad (1)$$

On the other hand, the NVT equilibrated configuration was also used as the starting structure (system S_4^{CG}) for a 3 ns NPT simulation at 300 K and 1 bar, from which the isothermal compressibility was computed according to[43]

$$\kappa = \frac{\langle V^2\rangle - \langle V\rangle^2}{\langle V\rangle k_B T} \quad (2)$$

Radial distribution functions (CG ion−WT4) and an electrostatic potential profile (obtained in the same way as in the atomistic system) were calculated for systems S_5^{CG} and S_6^{CG} and compared with systems S_2^{AA} and S_3^{AA}, respectively, in order to assess the ability of the CG model to reproduce atomistic results.

Systems S_7^{CG} and S_8^{CG} were used to compute radial distribution functions (CG ion-WT4) using an ionic concentration of roughly 0.5 M, in order to compare them with experimental data.[30]

Bjerrum (λ_B) and Debye (κ^{-1}) lengths were calculated as

$$\lambda_B(\rho) = \frac{(1.0 \times 10^9)\beta\, e^2}{4\pi\varepsilon_0\, \varepsilon_r(\rho)} \quad (3)$$

$$\kappa^{-1}(\rho) = \left(\frac{2(1.0 \times 10^{-15})F^2\rho}{RT\varepsilon_0\, \varepsilon_r(\rho)}\right)^{-1/2} \quad (4)$$

where β is the thermal energy, $\varepsilon_0 = 8.85 \times 10^{-12}$ C^2 J^{-1}m^{-1}, $F = 96485.3399$ C mol^{-1}, and $R = 8.314472$ J mol^{-1} K^{-1}.

The dielectric constants of the electrolyte aqueous solutions (Na^+Cl^- and K^+Cl^-) at different concentrations (0.2 and 0.3 M) were obtained from simulations of systems S_{9-12}^{CG} (see Table 2).

The osmotic pressure measurement was based on the methodology presented by Roux and Luo[47] (systems $S_{13,14}^{CG}$). The idea behind it is to simulate an aqueous solution where the ions are restrained to stay only in one-half of the simulation box and from the force exerted by the restraints, calculate the osmotic pressure. To accomplish this, we used a restraining strategy previously developed in our group called BRIM[48] (see the Supporting Information for a more exhaustive explanation).

Finally, we performed a 4 μs unconstrained simulation (S_{15}^{CG}) of a CG version[24] of a double-stranded DNA using the Drew−Dickerson sequence 5′-d(CGCGAATTCGCG)-3′ in an octahedron box filled with WT4 and CG ions (see Table 2). Global DNA structural behavior, DNA hydration, and specific DNA−water and DNA−ion interactions were evaluated. Helical parameters for DNA were computed using the Curves+ software.[49] The cation-induced narrowing of the minor groove was studied. Such structural changes were estimated from the average interphosphate distance between opposite strands measured for the following pairs: {(*5*, 24), (*6*, 23), (*7*, 22), (*8*, 21), (9, *20*), (10, *19*), (11, *18*), (12, *17*)} (italics indicate the residue numbers at the AT track). Cations were considered to be bound to the minor groove if their distance to the phosphate groups of both opposite strands was below 0.5 nm.

Results

In the following paragraphs, we describe the performance of the WT4 model to reproduce some common parameters of pure water. Comparisons are made, whenever possible, against experimental data. However, some of the calculated properties are also confronted with the results obtained from popular atomistic water models just to provide a reference frame for our results against well established AA models used by the broad scientific community. Subsequently, we analyze the solvation structure of simple electrolyte representations. Finally, to provide an example of application to a biologically relevant system, we briefly present a simulation of a CG DNA double helix in the presence of explicit solvent and mixed salts. A more detailed study on different properties of DNA (flexibility, breathing, DNA−solvent interactions on the multi-microsecond time scale, etc.) will be published elsewhere.

WT4 in the Bulk. A characteristic feature of water is its intrinsic ordering. A good reproduction of the oxygen−oxygen radial distribution function is a common goal for most water models in atomistic detail. The shape of the radial distribution function (RDF) at points far from the first spheres of hydration may furnish an idea of the liquid character of the substance under study. While for a liquid the RDF is expected to converge to a unitary value after a certain point, repetitive behavior is indicative of a crystalline state.

Although the RDF obtained with our model retains some characteristic features of liquid water, comparison of the RDF

obtained for WT4 with other atomistic models reveals some dissimilarities. The most evident difference with respect to the RDF calculated for SPC or TIP3P simulations (systems S_1^{AA} and S_4^{AA}) is the complete lack of the first solvation peak. Owing to the size and topology of the beads, WT4 presents a void space from the center of each bead up to the distance corresponding to the second solvation shell of real water. In this sense, the WT4 representation can be considered a second shell solvation model. In fact, the position of the first maximum in WT4 corresponds to the second peak of atomic water[50] (Figure 2A). It is important to notice that the normalization to the bulk value and the more granular character of the CG model generates a difference in the relative heights of the probability distribution of WT4 with respect to real water. Furthermore, the harmonic bonds existing within the tetrahedron translate into an overestimation of the probability of finding the first neighbor in the WT4 solution. After this global maximum, the relatively large size of WT4 generates some residual ordering that extends up to ~1.2 nm. The radial distribution function converges to one (bulk density) beyond 1.3 nm.

An important property for models of liquid water is their capability to reproduce the correct water diffusion. Clearly, the diffusivity of the WT4 molecules is much lower than that of atomistic water. At 300 K, we obtained a value of 2.03×10^{-6} cm^2 s^{-1}. However, the displacement of a WT4 molecule implicitly represents the movement of the center of mass of ~11 water molecules. Taking into account that the average mean squared displacement of the center of mass of N molecules is N times slower than the average mean squared displacement of N molecules diffusing separately,[53,54] we can conclude that the self-diffusion coefficient for the water molecules represented by the CG model at room temperature is 2.23×10^{-5} cm^2 s^{-1}, which is in good agreement with the experimental value (Table 3).

The WT4 model includes the explicit treatment of the electrostatic interactions as each bead carries a point charge (Table 1). This gives rise to a dielectric permittivity without imposing a continuum dielectric medium. The dielectric permittivity simulated by WT4 is 110.[55] Although this value is nearly 30% higher than that of real water, it must be noticed that this has been a problematic point even for more sophisticated atomistic models of water, and values ranging from 53[56] to 116[57] have been reported.

An important issue regards the long-range ordering of the WT4 molecules. In fact, some CG models for water present a freezing point very close to room temperature.[41] Therefore, we sought to perform a temperature scan over a range from 278 to 328 K. This range of temperatures covers most of the potential and biologically relevant applications of the model. Calculation of the RDF along the studied temperature range suggests that WT4 retains its liquid character, as no significant changes are found between 278 and 328 K (Figure 2A, inset).

The density of the WT4 model was set to match the value of pure water at 300 K. However, a reasonably good reproduction of the variations of the density versus temperature is also desirable. From the qualitative point of view, we obtained the expected reduction of the density with the

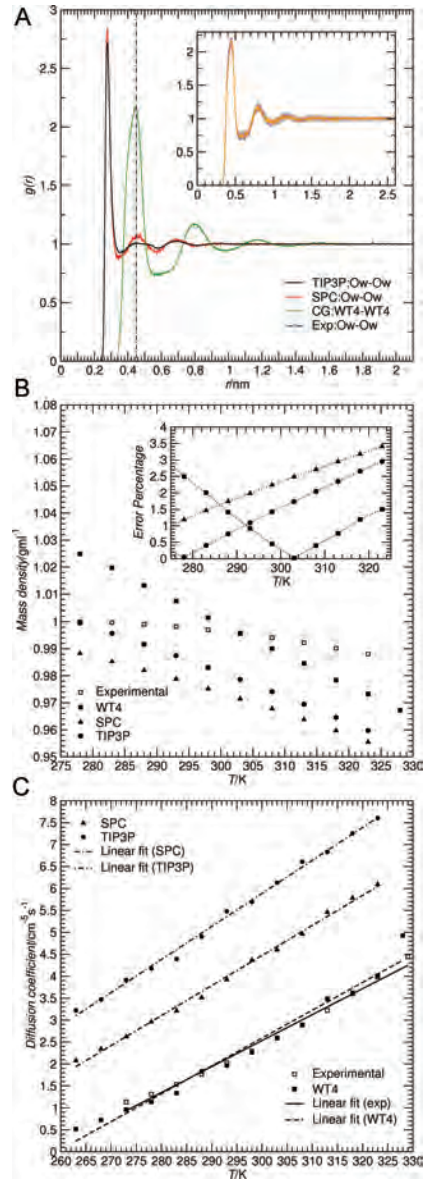

Figure 2. Bulk properties of WT4. (A) RDF calculated over all of the WT4 beads at room temperature from system S_1^{CG} (green line). Comparison is made with the oxygen−oxygen RDF calculated from TIP3P and SPC atomistic simulations as obtained from systems S_1^{AA} and S_4^{AA} at 298 K (black line and red line, respectively). The position of the second solvation peak obtained from experiments[50] is also shown (vertical, dot-dashed line). The inset shows the behavior of the RDF upon temperature variations (system S_2^{CG}) in the range from 278 to 328 K. No significant changes are observed. (B) The variation of the CG water mass density with the temperature (filled squares) calculated from system S_2^{CG} as compared with experimental data (empty squares)[51] and simulations of SPC (triangles) and TIP3P (circles) systems (S_1^{AA} and S_4^{AA}, respectively). The inset shows the relative error of the WT4, SPC, and TIP3P models compared to the experimental data. (C) The dependence of the diffusion coefficient on temperature is compared between the WT4, SPC, and TIP3P models (S_2^{CG} (filled squares), S_1^{AA} (triangles), and S_4^{AA} (circles), respectively) and experimental data[52] (empty squares). All four profiles present an almost linear trend, as revealed by the corresponding linear fits.

Table 3. Bulk Water Properties at Room Conditions for Atomistic Water Three-Point Models (SPC and TIP3P), WT4, and Experimental Data

	dielectric constant	diffusion coefficient (10^{-5} cm^2 s^{-1})	expansion coefficient (10^{-4} K^{-1})	mass density (g mL^{-1})	number density[a] ($\times$ 10^{22} mL^{-1})	surface tension (mN m^{-1})	isothermal compressibility (GPa^{-1})
WT4	110	2.23	11.6	1.0001	0.3	17	2.43
SPC	65[58]	3.85[59]	7.3[60]	0.9705[61]	3.2	53.4[62]	0.53[63]
TIP3P	82[56]	5.19[59]	9.2[64]	1.002[64]	3.4	49.5[62]	0.58[63]
Exp.	78.4[65]	2.27[66]	2.53[51]	0.9970[51]	3.3	71.2[67]	0.46[68]

[a] Calculated from the corresponding mass density, considering the molar mass of water (18 g mol^{-1}) and WT4 (200 g mol^{-1}). Accordingly, the number density for the atomistic models and real water corresponds to the number of water molecules per milliliter, while for WT4 it corresponds to the number of WT4 molecules (~11 water molecules) per milliliter.

temperature and an almost perfectly linear behavior of the system's density against temperature in the explored range (Figure 2B). Although the functional dependence of real water against temperature is certainly not linear, it is a good approximation within the temperature range chosen. In fact, the relative error of the WT4 density with respect to that of the real water in this temperature window remains always below 3%, with the higher deviations near the critical point of real water (Figure 2B). This behavior is comparable with those of the SPC and TIP3P atomistic models (Figure 2B).

Following the volume changes upon thermal variations at constant pressure allows also for the calculation of the isobaric expansion coefficient of our model. We obtain an overestimation of this quantity at 298 K (Table 3). The expansion coefficient of WT4 gives a value of 11.6×10^{-4} K^{-1} as compared with the experimental value of 2.53×10^{-4} K^{-1}.[51] Although overestimated, it is comparable with the values reported for widely used three-point water models (Table 3).

Another frequently calculated property for CG models is the surface tension. In our case, we obtained a value of 17 mN m^{-1}, which is nearly 4 times smaller than the experimental value. Similarly, we found a 5 fold higher isothermal compressibility as compared with the experimental value (Table 3). These discrepancies are very frequently found in CG models that lump a number of water molecules into one single entity.[42] The origin of this effect may be the loss of fully atomic interactions that decrease the cohesive forces and increase the granularity of the system.

A more stringent test for our representation comes from the calculation of the diffusion coefficient. Clearly, a rise in the diffusion must occur upon heating. Experimental data indicates that pure water experiences a nearly linear increase in the diffusion coefficient between 278 and 328 K. The model shows the correct dependence of the diffusion coefficient on temperature. Indeed, it shows good agreement with the experimental behavior within the explored range (Figure 2C).

Taking into account the above results, the range of validity of the model may be delimited by the following considerations: the lower limit should not go below 278 K. Applications at lower temperatures are strongly discouraged since ice formation implies quantum effects that can, obviously, not be achieved by simplified models. On the upper limit, the relative error in the renormalized diffusion coefficient arrives at ~11% at 328 K, suggesting that simulations at higher temperatures could require some reparameterization to keep the accuracy at acceptable levels.

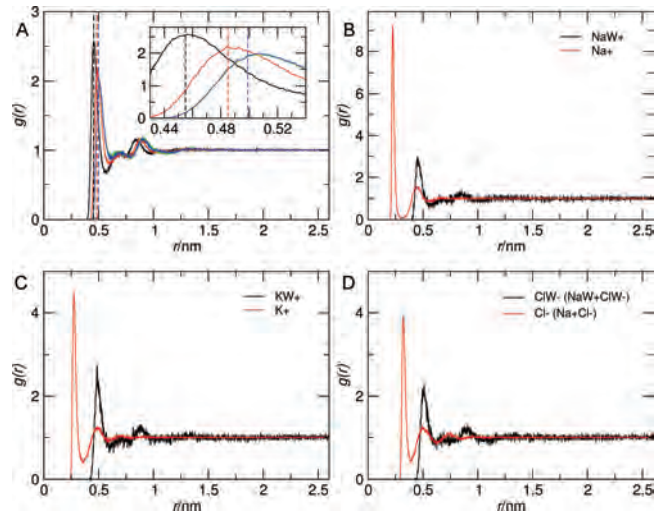

Figure 3. Ionic solvation. (A) The RDF of WT4 around CG electrolytes computed for systems S_7^{CG} and S_8^{CG} (NaW, black; KW, red; ClW, blue for NaW^+ClW^- and green for KW^+ClW^-). Vertical dashed lines indicate the position of the second solvation peak as determined from neutron scattering experiments.[30] The inset shows a closeup on the region between 0.43 and 0.54 nm allowing for a more precise comparison. (B, C, and D) Comparison between RDFs obtained from atomistic and CG simulations (systems S_2^{AA}, S_3^{AA}, S_5^{CG}, and S_6^{CG}). The plot corresponding to the solvation structure around chlorine ions in the presence of potassium is similar to that shown for the case of sodium. It is omitted for brevity.

Ionic Solvation. The characteristics of the WT4 model open the possibility to study the solvation properties of systems in which electrostatics are dominant. In this context, we developed the CG parameters of three simple electrolytes: Na^+, K^+, and Cl^-. Since we can imagine WT4 as a second solvation shell model, we represent ions together with their first sphere of hydration. Aimed at exploring the solvation structure generated by the WT4 model on the CG ions, simulations were conducted at roughly similar ionic concentrations to those reported in neutron diffraction experiments.[30] As depicted in Figure 3A, there is good correspondence, especially for the cations, between the first solvation maximum found for WT4 and the second hydration shell estimated from the experimental data.[30] A second solvation peak is found at nearly 0.9 nm, which has to be considered mainly as an artifact of the geometry of the model since the beads located in the last peak are harmonically linked to those

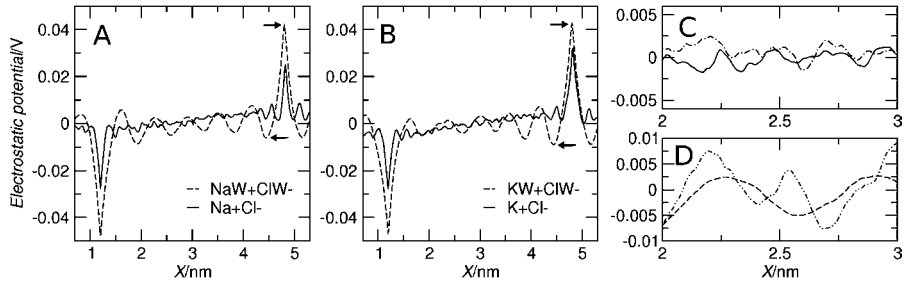

Figure 4. Profiles of electrostatic potential. (A) Electrostatic potential calculated along the line connecting the ionic pairs Na^+Cl^- (filled line, S_2^{AA}) and NaW^+ClW^- (dashed line, S_5^{CG}). Arrows indicate the points where the differences in the electrostatic potential were calculated. (B) Same as A for systems K^+Cl^- (S_3^{AA}) and KW^+ClW^- (S_6^{CG}). (C) Comparison of the electrostatic potential between the central portion of panel A (filled line, S_2^{AA}) against the analogous quantity calculated along a box containing pure SPC water (dot-dashed line, S_1^{AA}). (D) Same as C but for the CG systems (dashed line, S_5^{CG}, and double-dot-dashed line, S_1^{CG}).

of the first. After that point, the RDF converges to the bulk value in all cases.

Unfortunately, experimental data for ionic solvation is only available at high electrolyte concentrations. To explore lower (and more physiological) concentrations for which no experimental data are available, we tried a comparison with atomistic simulations confronting systems S_2^{AA} and S_3^{AA} with systems S_5^{CG} and S_6^{CG}, respectively, both having an ionic concentration of 0.01 M (Figures 3B, C, and D).

In close analogy with the case of pure WT4, the RDF of WT4 around CG ions shows a complete lack of the first solvation shell. A good reproduction of the position of the second solvation peak is observed, confirming the behavior of WT4 as a second solvation shell solvent. As expected, WT4 is not able to reproduce the third solvation shell. The relevance of this inaccuracy is, however, uncertain and could only be relevant in the case of chlorine ions, where such a shell is slightly more pronounced.[30]

Electrostatic Potential. Having analyzed the hydration structure of simple electrolytes, we turned our attention to the profiles of electrostatic potential and the screening properties. This was done by comparing the results of systems S_2^{AA} and S_3^{AA} against those of S_5^{CG} and S_6^{CG}, respectively. These systems consist of an ionic pair of Na^+Cl^- (or K^+Cl^-) kept at a fixed position during the simulation. The separation between both ions was 3.6 nm. Atomistic ionic pairs were immersed in a computational box containing an equivalent number of water molecules. This setup allowed us to compare under similar conditions atomistic and CG simulations as well as the behavior of the different ionic species. In order to assign the proper weight to the perturbations introduced by the electrolytes, we also made comparisons with the fluctuations produced by pure solvent (atomistic and CG) in the profiles of the electrostatic potential. In this way, it is possible to separate the observed features into two components: intrinsic bulk fluctuations and ionic perturbations. Furthermore, this approach gives an idea about the relaxation of the ionic potential at increasing distances from the ion and compares it with pure water and electrolyte solution.

A comparative view of the atomistic versus CG simulation can be acquired from Figure 4A. The first noticeable difference regards the height of the peaks centered on the positions of the ions. Owing to its smaller size, the SPC waters can get closer to the atomistic ion generating a more pronounced electrostatic screening. In the CG counterpart, the corresponding first solvation shell, which is implicit in the NaW^+ and ClW^- ions, only serves to create a void space without screening properties. This translates to a higher electrostatic potential induced by the CG ion. The implicit consideration of the first solvation shell in the CG ions implies that the first minimum observed in the atomistic system is absent in the CG system (Figure 4A). Furthermore, the position of the first minimum observed in the CG simulation (second solvation shell) roughly corresponds to the position of the second minimum around the ion in the atomistic system. Clearly, this effect derives from the solvation structure around the electrolytes; i.e., the first and second minima around the position of the ions (both, Na^+ and Cl^-) shown in Figure 4A correspond to the position of the oxygen atom in the first and second solvation shells shown in Figure 3B and D. Similar features are observed for the cases of K^+Cl^- and KW^+ClW^- ionic pairs (Figure 4B).

The distinctive characteristics of both cations evidenced by the solvent organization around NaW^+ and KW^+ (Figure 3A) can also be obtained from the calculation of the difference in electrostatic potential measured at the position of the cation with respect to that of its first minima (Figure 4A,B). This difference was about 10% higher for the case of KW^+ with respect to NaW^+, in qualitatively good agreement with the ~25% obtained from the atomistic case. This behavior may reflect the fact that water around potassium is bound in a more disorderly fashion than around sodium,[29] probably generating a less marked electrostatic screening in the case of potassium.

As seen from Figure 4A and B, the CG scheme presents higher fluctuations in the potential than the atomistic system. Aimed at excluding the possibility of a spurious ordering of WT4 molecules around the electrolytes, we compared the perturbations in the electrostatic potential introduced by the ions against those observed for pure solvent (both, atomistic and CG). This was assessed by computing the electrostatic potential along an arbitrary axis in two simulation boxes containing pure SPC and WT4 (systems S_1^{AA} and S_1^{CG},

Table 4. Thermodynamic Properties of Electrolyte Solutions

		Bjerrum length (nm)		Debye length (nm)		osmotic pressure[a] (bar)
ρ (molarity)		0.2 M	0.3 M	0.2 M	0.3 M	0.5 M
NaW^+ClW^-/WT4		0.57	0.61	0.76	0.6	35 (s.d. 15)
KW^+ClW^-/WT4		0.55	0.61	0.78	0.6	33 (s.d. 16)
Exp.[b]	NaCl	0.75	0.77	0.66	0.53	~25 (taken from Roux and Luo[47])
	KCl	0.74	0.76	0.67	0.54	

[a] The value obtained in the atomistic simulations using CHARMM PARAM 27 was 37 (s.d. 9) bar. [b] The function $\varepsilon_r(\rho) = \varepsilon(0)/(1 + A\rho)$ (NaCl, $A = 0.27$; KCl, $A = 0.24$), which results from fitting to experimentally obtained dielectric constants,[69] was used to estimate $\varepsilon_r(\rho)$ at the desired concentration, which is necessary for the computation of both Bjerrum and Debye lengths.

respectively). Superposition of both profiles (Figure 4C and D) suggests that both pure water systems show important fluctuations in the electrostatic potential of nearly the same magnitude as those observed in the region between the ions in the ionic solution. This indicates that the perturbations observed in those regions are not an effect induced by the ions but correspond to variations in the electrostatic potential, which are intrinsic to the pure solution. According to this, the difference in the amplitude of the fluctuations observed between the atomistic and CG models (Figure 4A and B) are explained by the augmented granularity of the CG model. An estimation of such a difference is obtained from the approximate amplitudes observed in both atomistic (~0.002 V) and CG (~0.018 V) simulations. This indicates that the oscillations in the CG system have amplitudes nearly 1 order of magnitude higher than the ones in the atomistic system.

Bulk Electrolytic Properties. The vast majority of empirical parametrizations for single ions are typically developed to fit single ion properties, such as those examined in the previous sections. In order to complement the structural description of the CG aqueous solutions we studied some thermodynamic properties regarding ion−ion interactions: in particular, the Bjerrum and Debye lengths. The first represents the separation between two elementary charges at which the electrostatic interaction is comparable in magnitude to the thermal energy, and the second provides information regarding the distance at which the electrostatic potential of one ion is screened by the ionic strength of the surrounding medium. From the qualitative point of view, we retrieved the correct tendency in Bjerrum and Debye lengths upon changes in the ionic concentration (Table 4). Calculation of the Bjerrum and Debye lengths at 0.2 and 0.3 M gave values within a maximum error of 13% with respect to experimental values (Table 4). We obtained an underestimation of the Bjerrum length and, correspondingly, an overestimation of the Debye length, which is indicative of a slightly higher global electrostatic screening in the bulk solution, independent of the salt considered in the simulation (i.e., NaW^+ClW^- or KW^+ClW^-).

A direct measurement of the strength of the effective solvent-mediated interaction between ions is also very relevant, and it can be obtained from the osmotic pressure. For the case of NaW^+ClW^- at an ionic concentration of 0.5 M, we obtained a value of 35 bar (33 bar for KW^+ClW^-), which is essentially identical to that obtained by atomistic simulations using the CHARMM force field. Despite the large standard deviations, these values are in agreement with experimental reports (Table 4), suggesting a satisfactory balance in ion−ion and ion−WT4 interactions.

CG Solvation of Double-Stranded DNA. As a final example of application, we analyzed the explicit solvation of a dodecameric segment of double-stranded DNA. For this task, we used the already published CG scheme for simulating nucleic acids within the framework of the generalized Born model for implicit solvation.[24] In this contribution, the same system was simulated in the presence of explicit solvent and added salts. Both approaches furnish a similar picture of the structural and dynamical behavior of the double-helical segment of DNA with a maximum pairwise RMSD between both average structures of 0.25 nm. This is in agreement with the good reproduction of helical parameters obtained upon backmapping from the DNA simulation in explicit CG solvent (Figure S2, Supporting Information). Furthermore, the superposition of the covariance matrices calculated along the MD trajectories of the Drew−Dickerson dodecamer performed using implicit and explicit solvation gives an identity of 84%. This strongly suggests that both approaches sample nearly equivalent conformational spaces.

During the dynamics in the presence of explicit CG solvent, the global structure of the DNA dodecamer was fairly well conserved with an average RMSD of 0.25 nm from the starting (canonical) conformer. This can be inferred from the good superposition of snapshots taken at different times of the simulation (Figure 5A). Moreover, a good agreement is also obtained at the atomistic level upon backmapping. The all atoms RMSD of those shapshots compared with the X-ray structure 1BNA resulted in values of 0.35 nm (blue), 0.39 nm (green), and 0.34 nm (orange) (Figure S3, Supporting Information).

The WT4 molecules and cations closely interact with the CG nucleobases. It can be observed that the ordering of the WT4 molecules around the DNA qualitatively resembles the hydration features encountered in atomistic systems at both experimental and theoretical levels.[70−75] Conical arrangements of WT4 beads form around the phosphate groups (Figure 5B). The molecules of WT4 acquire an orientation guided by the electrostatic attraction between the positive (hydrogen-like) beads and the negatively charged phosphate superatoms. This results in the formation of structures alike to hydration cones (Figure 5B). In this kind of solvent arrangement around the backbone, WT4 molecules can be replaced by cations from the solution (Figure 5B) as observed experimentally.[76] Furthermore, ions can also remain transiently bound to the DNA visiting different positions within the minor groove. Extended hydration of the major groove and the formation of hydration spines in the minor groove are also observed, as illustrated in Figure 5C. A comprehensive picture of the hydration structure can be obtained from the WT4 beads' occupancy density map projected in the plane perpendicular to the DNA axis placed at the AT step (Figure 5D).

A more quantitative view of the solute/solvent interaction can be obtained from the cumulative RDF of the different species around the phosphate superatoms (Figure 5E). The

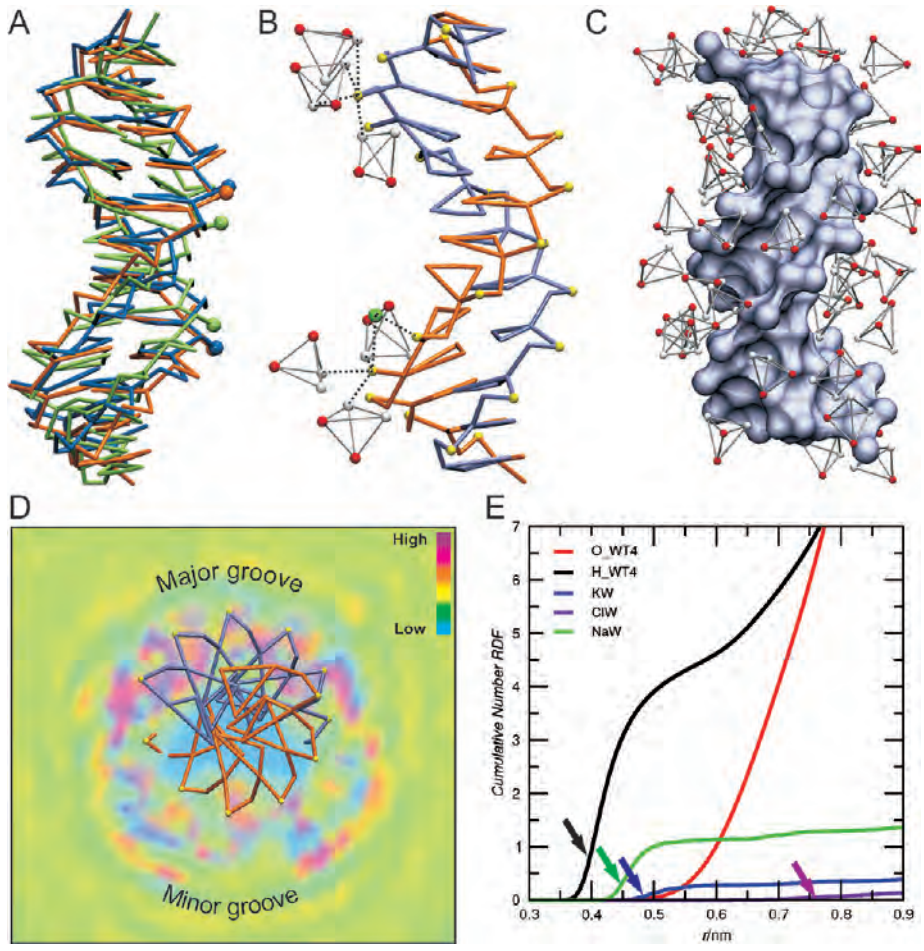

Figure 5. DNA and solvation structure. (A) Superposition of the DNA conformers taken from the first (blue), middle (green), and last (orange) frame of the simulation of system S_{15}^{CG}. Spheres indicate a pair of phosphate superatoms from opposite strands, which are highlighted in order to show the minor groove narrowing. An atomistic view of this superposition obtained from backmapped CG coordinates can be seen in Figure S6, Supporting Information. (B) WT4 and NaW specific interaction with the phosphate groups taken from a representative MD snapshot. The dashed lines highlight the conical arrangement of WT4 beads around phosphates (top) and the competition for the phosphate groups by WT4 and NaW (bottom). (C) WT4 solvation in the major and minor grooves from a random frame. The extensive hydration of the major groove and spines of hydration within the minor groove are evident. (D) WT4 occupancy density map projected onto a plane orthogonal to the DNA axis, located in the central AT step. The color scale represents the occupancy level, with a color range from cyan (low occupancy) to purple (high occupancy). Differences in major and minor groove are plain. Notice also the more punctuated location of WT4 within the minor groove indicative of solvation spines. (E) Cumulative number (integral of the RDFs) of negative and positive WT4 beads (red and black, respectively), NaW^+ (green), KW^+ (blue), and ClW^- (violet), with respect to the phosphate groups. Arrows indicate inflection points, which correspond to the first maxima of each RDF.

directionality in the WT4− phosphate interaction is evident from the right shift observed in the integral of the RDF corresponding to the oxygen-like beads' position with respect to that of the hydrogen-like beads (compare red and black lines in Figure 5E). The position of the first WT4 solvation shell forming conical structures lies at 0.4 nm from the phosphate superatom. This distance is in good agreement with the 0.38 nm found in atomistic simulations[24] and is comparable to the minimum distance of 0.32 observed in X-ray structures.[77]

Our model can also take into account the specificity in the DNA−cation interactions. As expected, sodium ions are more prone than potassium to interact with the solute. As mentioned above, sodium is frequently found in the close neighborhood of the phosphate moieties and even within the minor groove.[76,78,79] The closest sodium shell is localized at 0.45 nm from the phosphate, as compared with the 0.5 nm found for the bulkier potassium. In contrast, the radial distribution of the chlorine ions is much more shifted to the right with a first peak at 0.76 nm (Figure 5E).

The fraction of DNA charge neutralized within a cylinder of 0.9 nm from the exterior surface of the double-stranded helix is 0.75. This is in good agreement with the fraction of condensed counterions calculated within the condensation volume using Manning's counterion condensation theory for polyelectrolytes.[80] Moreover, this number is comparable with a fraction of 0.76 obtained by previous atomistic simulations using the same DNA sequence.[81] Among the fraction of condensed counterions, 76% corresponds to sodium and 24% to potassium; this is in qualitative agreement with a series of experimental and theoretical studies (see Savelyev and Papoian[82] and references therein).

While the global distribution of cations around the DNA contributes to the stability of the double helix, the specific

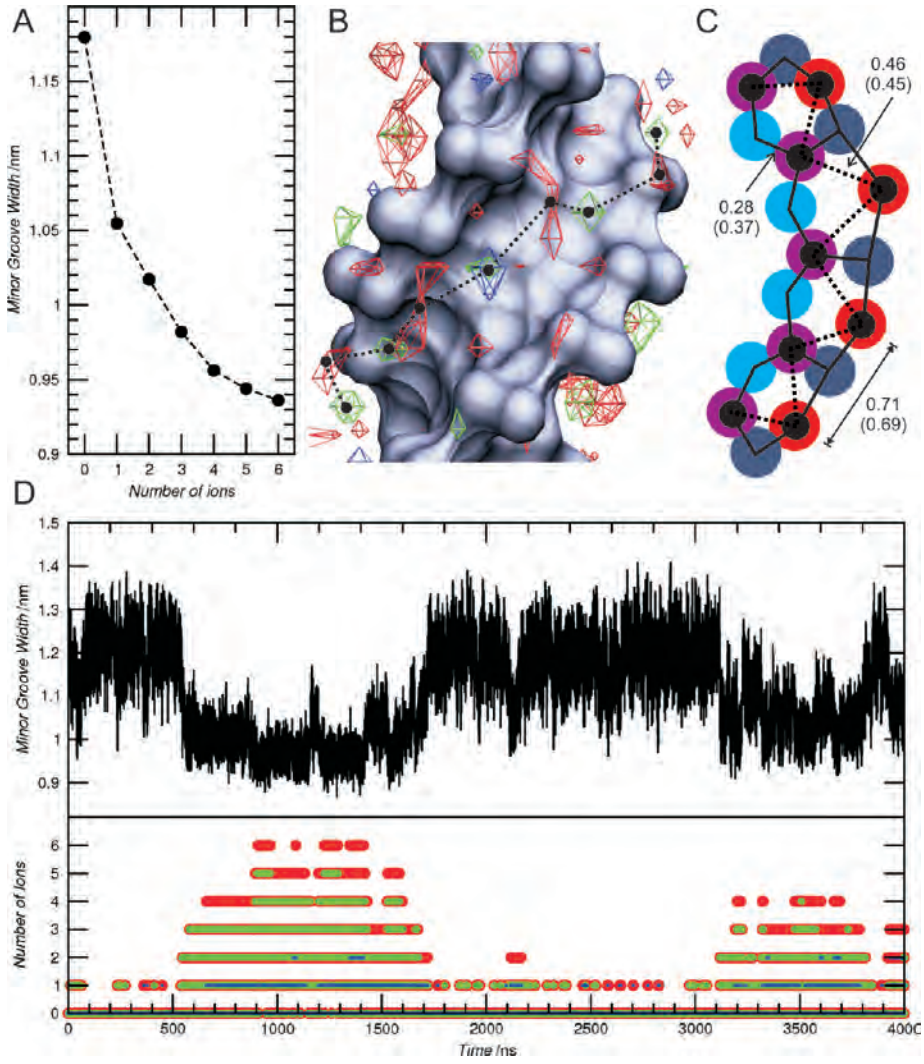

Figure 6. Binding of cations within the minor groove. (A) The minor groove width averaged over all frames with an equal number of bound cations is plotted against the number of bound ions (according to the criteria explained in the Methods section). (B) WT4 (red), NaW^+ (green), and KW^+ (blue) occupancy isosurfaces located in the minor groove of the AT track. Dashed path connecting black points indicate the zig-zag motif formed by the cations and WT4 beads in the minor groove. (C) Scheme showing the superimposition of the zig-zag motif (black circles connected by dashed line) observed in the CG simulation over the fused hexagon motif (continuous line) formed by the solvent sites (cyan, violet, gray, and orange circles) experimentally observed. These sites can by occupied by both water or cations.[77] The distances between corresponding solvation sites in the fused hexagon motif are shown and compared to the corresponding ones in the zig-zag motif (parentheses). (D) Minor groove width (top) and number of bound cations plotted against time (bottom). The number of cations is shown as the number of NaW^+ (green), number of KW^+ (blue), and total number of cations (sum of the number of NaW^+ and number of KW^+, in red).

interaction of cations with DNA has been related to local structural distortions. In particular, the binding of sodium ions within the minor groove has been proposed to mediate a narrowing in the minor groove.[83] In agreement with this proposal, we observed a clear correlation between the width of the minor groove and the binding of cations. Moreover, there seems to be a cumulative effect between these two events; i.e, a higher number of bound ions induces a more pronounced narrowing. This is clear from a measure of the average width of the minor groove with respect to the total number of bound ions (Figure 6A). The binding of one single ion is enough to induce a sensible change in the minor groove. Upon the successive incorporation of ions, the narrowing becomes more marked, reaching a minimum when six ions are concomitantly bound. Experimental studies on the same dodecamer also reveal a high occupancy of cations in the minor groove, leading to its narrowing.[78] Furthermore, a highly ordered structure is formed when cations and water interact with the AT track of the DNA. Such a structure is organized in four layers of solvent sites and resembles a series of fused hexagonal motifs.[78] Figure 6B shows the 3D occupancy map of WT4 and cations around the minor groove of the AT track. This map reveals sites highly occupied by WT4 (red wire mesh), NaW^+ (green wire mesh), and KW^+ (blue wire mesh) that resembles a zig-zag structure (dashed path connecting black points in Figure 6B). When such a zig-zag structure is superimposed onto the experimentally observed fused hexagon motif, good agreement is obtained for the second and fourth solvent-site layers, as confirmed by the inter solvent-site distances (Figure 6C).

The minor groove narrowing process appears to take place on two different time scales. The first is related to the binding

of one or two ions for up to a few dozen nanoseconds, while the second corresponds to the simultaneous binding of three to six ions for a period of nearly 1 μs (Figure 6D). This last induces a more marked and persistent but always reversible structural distortion with an average minor groove width of 0.98 nm (three bound cations) to 0.94 nm (six bound cations). The magnitude of this DNA distortion is in very good agreement with the average value of 0.96 nm obtained experimentally.[79]

It is worth noticing that temporal scales for sodium binding are coincident with the faster events found in this study have been reported for MD simulations at the atomistic level.[83−85] Unfortunately, the longest atomistic simulation reported in this system was carried out for 1.2 μs.[83] Although only nanosecond binding events were reported in that work, the agreement of the position of the binding sites and DNA distortion with X-ray data[78,79] may allow for speculation that a lack of longer binding events in the atomistic simulation could be related to insufficient sampling. Clearly, longer simulation times that go beyond the introductory scope of this paper would be needed to properly sample these long lasting events. This issue will be addressed in a forthcoming publication.

There is a marked selectivity for sodium against potassium. Indeed, while the simultaneous binding of more than two sodium ions is very frequent, only two potassium ions were present within the minor groove simultaneously, and this rather rare event was detected only five times in the 4 μs trajectory (Figure 6D and Figure S4, Supporting Information).

Finally, to complete the picture regarding the ionic structure around DNA, we analyzed the ionic distribution at longer distances from the double helix. This was done by calculating the number density of the three types of ions present in the system at increasing distances from DNA. In good agreement with the prediction from Poisson−Boltzmann theory,[86] the amount of electrolytes along a direction perpendicular to the DNA principal axis follows an exponential decay (Figure S5, Supporting Information).

Discussion and Conclusions

In this work, we have presented a model for simulating water at a coarse grain level. The WT4 model presented here is based on the transient tetrahedral structure adopted by water molecules in solution, preserving the molecular characteristics of the atomistic liquid. Due to the large number and heterogeneity of the CG models proposed in the literature, it is difficult to establish a fair comparison in terms of a computational speedup obtained with WT4. However, a comparison is more straightforward if we restrict it to the simplest models that condense three or four water molecules into a single bead.[53,54,87] This implies a coarse graining factor from 9 to 12, as compared to the value of ∼8 obtained for WT4. In addition to a similar coarse graining factor, our model offers some advantages, like the capacity to interact via explicit short- and long-range electrostatic interactions, and a dielectric permittivity. This grants the model the ability to reproduce some of the characteristic properties of water and electrolytic solutions.

The bead's masses were assigned to fit the water density at 300 K. Although this may raise some concerns about the suitability of the model at different temperatures, the relative error for the WT4 density with respect to the experimental determination of pure water remains below 3% in the range of 278 to 328 K (Figure 2B).

A strong assumption of the model is the fact that the existences of these five-member water clusters are supposed to be permanent, while their average lifetime in real water is on the order of picoseconds. This defect is partially compensated by setting a loose harmonic constraint between the beads of our representation. This allows for bond stretching variations of about 10% in the bond lengths, conferring a large plasticity to the WT4 molecules and the possibility of adapting their conformation according to its molecular environment.

The use of the WT4 model to solvate simple ions reproduces their hydration structure and some thermodynamic properties such as osmotic pressure, which is often considered a quality gauge of the parametrization.

We notice that important properties such as the isothermal compressibility and surface tension are poorly described by WT4. This may be of particular relevance in the study of self-assembly phenomena, and for such treatment special caution is advised. Despite this deviation from ideal behavior, the description of the double-stranded DNA segment does not seem to be compromised. This suggests that the long−medium range screening properties of the solvation model are suitable for overcoming the strong electrostatic repulsion generated by the negatively charged phosphate groups of DNA. In fact, the addition of explicit solvation and different ionic species highly enhanced the description of the DNA dynamics, allowing, for instance, the reproduction of the cation-mediated narrowing of the minor groove that could not be studied within the implicit solvation approach. While the implicit solvation approach can provide a good and faster description of sequence dependent effects on the structural and dynamical stability, inclusion of the explicit solvent can allow for the study, for instance, of the influence of intrinsic versus extrinsic sources of DNA flexibility, solvent mediated effects, ionic specificity, etc. Furthermore, the use of periodic boundary conditions and explicit electrostatics permits a more realistic consideration of long-range effects.

WT4 together with the CG electrolyte model represent correctly the gross solvation structure around DNA, as noted by the percentage of DNA charge neutralized at 0.9 nm that closely resembles that of atomistic simulations and that predicted by counterion condensation theory. Moreover, DNA hydration features like the extensive major groove hydration, minor groove hydration spines, and conical arrangement around phosphate groups that resembles the hydration cones observed in atomistic simulations and experimental data are well reproduced. It is important to note in this context that the development of interaction parameters has always been carried out within the philosophy of fitting structural properties of water, ionic solvation, and DNA. In this respect, we first developed the representation for WT4 in the bulk and then added the description of simple CG

electrolytes. Finally, the existing DNA parameters for implicit solvation were slightly modified to further refine its structural description when embedded in explicit solvent. In this sense, good agreement with experimental determinations can be considered emerging properties of the model because no specific fittings of cross interaction potentials have been performed.

The simulation scheme presented here allows for runing at a rate of ~1 μs per day (S_{15}^{CG}) on a dual quad core PC (Intel Xeon 2.66 GHz). This performance along with the nearly atomic resolution achievable upon backmapping of the coordinates in our DNA model[24] make the millisecond time scale reachable. This would effectively bridge the gap between the time scales feasible to MD and those that are biologically relevant.

Finally, we would like to stress the point that the model presented here computes all of the interactions using a typical Hamiltonian function, avoiding *ad hoc* code modifications/recompilations. Topologies, interaction parameters, and coordinate files for GROMACS implementation are available from the authors upon request.

Acknowledgment. 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) and Grant FCE_60-2007. L.D. and M.R.M. are beneficiaries of the National Fellowship System of ANII.

Supporting Information Available: Mapping scheme between atomistic and CG model for DNA. Helical parameters of the backmapped trajectory. Table containing interaction parameters sets for DNA. Energy fluctuations analysis. Cations in the minor groove frequency. Supperposition of backmapped DNA structures. Details on osmotic pressure calculation. Long range ionic structure. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

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

(2) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008**, *321*, 798–800.

(3) 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.

(4) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694–698.

(5) Tanaka, S.; Scheraga, H. A. Statistical Mechanical Treatment of Protein Conformation. I. Conformational Properties of Amino Acids in Proteins. *Macromolecules* **1976**, *9*, 142–159.

(6) Yin, Y.; Arkhipov, A.; Schulten, K. Simulations of membrane tubulation by lattices of amphiphysin N-BAR domains. *Structure* **2009**, *17*, 882–892.

(7) Arkhipov, A.; Yin, Y.; Schulten, K. Four-scale description of membrane sculpting by BAR domains. *Biophys. J.* **2008**, *95*, 2806–2821.

(8) Wee, C. L.; Gavaghan, D.; Sansom, M. S. P. Interactions between a voltage sensor and a toxin via multiscale simulations. *Biophys. J.* **2010**, *98*, 1558–1565.

(9) Ayton, G. S.; Voth, G. A. Hybrid coarse-graining approach for lipid bilayers at large length and time scales. *J. Phys. Chem. B* **2009**, *113*, 4413–4424.

(10) Treptow, W.; Marrink, S.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.

(11) Yefimov, S.; van der Giessen, E.; Onck, P. R.; Marrink, S. J. Mechanosensitive membrane channels in action. *Biophys. J.* **2008**, *94*, 2994–3002.

(12) Periole, X.; Huber, T.; Marrink, S.; Sakmar, T. P. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.* **2007**, *129*, 10126–10132.

(13) Durrieu, M.; Bond, P. J.; Sansom, M. S. P.; Lavery, R.; Baaden, M. Coarse-grain simulations of the r-snare fusion protein in its membrane environment detect long-lived conformational sub-states. *Chem. Phys. Chem.* **2009**, *10*, 1548–1552.

(14) Srinivas, G.; Discher, D.; Klein, M. Self-assembly and properties of diblock copolymers by coarse-grain molecular dynamics. *Nature* **2004**, *3*, 638–644.

(15) Nielsen, S.; Lopez, C.; Srinivas, G.; Klein, M. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

(16) Arkhipov, A.; Freddolino, P. L.; Schulten, K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* **2006**, *14*, 1767–1777.

(17) Srinivas, G.; Klein, M. Molecular dynamics simulations of self-assembly and nanotube formation by amphiphilic molecules in aqueous solution: a coarse-grain approach. *Nanotechnology* **2007**, 18.

(18) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophys. J.* **2007**, *92*, 4289–4303.

(19) Voth, G. A. *Coarse-graining of condensed phase and biomolecular systems*, 1st ed.; Taylor & Francis Group: New York, 2009; pp 1−455.

(20) DeMille, R. C.; Molinero, V. Coarse-grained ions without charges: reproducing the solvation structure of NaCl in water using short-ranged potentials. *J. Chem. Phys.* **2009**, *131*, 034107.

(21) Molinero, V.; Moore, E. B. Water modeled as an intermediate element between carbon and silicon. *J. Phys. Chem. B* **2009**, *113*, 4008–4016.

(22) Savelyev, A.; Papoian, G. A. Molecular renormalization group coarse-graining of electrolyte solutions: application to aqueous NaCl and KCl. *J. Phys. Chem. B* **2009**, *113*, 7785–7793.

(23) Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.* **2010**, *6*, e1000810.

(24) Dans, P.; Zeida, A.; Machado, M.; Pantano, S. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.* **2010**, *6*, 1711–1725.

(25) Head-Gordon, T.; Hura, G. Water structure from scattering experiments and simulation. *Chem. Rev.* **2002**, *102*, 2651–2670.

(26) Narten, A. H.; Danford, M. D.; Levy, H. A. X-ray diffraction study of liquid water in the temperature range 4−200°C. *Discuss. Faraday Soc.* **1967**, *43*, 97–107.

(27) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331−342.

(28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(29) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. Hydration of sodium, potassium, and chloride ions in solution and the concept of structure maker/breaker. *J. Phys. Chem. B* **2007**, *111*, 13570–13577.

(30) Mancinelli, R.; Botti, A.; Bruni, F.; Ricci, M. A.; Soper, A. K. Perturbation of water structure due to monovalent ions in solution. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2959–2967.

(31) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. Gromacs: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(32) Lindahl, E.; Hess, B.; van der Spoel, D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001**, *7*, 306–317.

(33) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. Gromacs: fast, flexible and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(34) Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; van Drunen, R.; van der Spoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R. *Gromacs: A parallel computer for molecular dynamics simulations*; de Groot, R. A., Nadrchal, J., Eds.; World Scientific: Singapore, 1993.

(35) Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255–268.

(36) Hoover, W. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(37) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(38) Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

(39) Darden, T.; 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.

(40) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. A smooth particle mesh ewald potential. *J. Chem. Phys.* **1995**, *103*, 8577–8592.

(41) Winger, M.; Trzesniak, D.; Baron, R.; van Gunsteren, W. F. On using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1934–1941.

(42) He, X.; Shinoda, W.; DeVane, R.; Klein, M. L. Exploring the utility of coarse-grained water models for computational studies of interfacial systems. *Mol. Phys.* **2010**, *108*, 2007–2020.

(43) Herrero, C. P. Compressibility of solid helium. *J. Phys.: Condens. Matter* **2008**, *20*, 295230.

(44) van Buuren, A. R.; Marrink, S. J.; Berendsen, H. J. C. A molecular dynamics study of the decane/water interface. *J. Phys. Chem.* **1993**, *97*, 9206–9212.

(45) Mark, A. E.; van Helden, S. P.; Smith, P. E.; Janssen, L. H. M.; van Gunsteren, W. F. Convergence properties of free energy calculations: alpha-cyclodextrin complexes as a case study. *J. Am. Chem. Soc.* **1994**, *116*, 6293–6302.

(46) Beglov, D.; Roux, B. Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *J. Chem. Phys.* **1994**, *100*, 9050–9063.

(47) Luo, Y.; Roux, B. Simulation of osmotic pressure in concentrated aqueous salt solutions. *J. Phys. Chem. Lett.* **2010**, *1*, 183–189.

(48) Herrera, E. F.; Pantano, S. Salt induced asymmetry in membrane simulations by partial restriction of ionic motion. *J. Chem. Phys.* **2009**, *130*, 195105–195114.

(49) Lavery, R.; Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **1988**, *6*, 63–91.

(50) Soper, A. K. The radial distribution functions of water and ice from 673 K and at pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121–137.

(51) Kell, G. S. Density, thermal expansivity, and compressibility of liquid water from 0° to 150°C: correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *J. Chem. Eng. Data* **1975**, *20*, 97–105.

(52) Holz, M.; Heil, S. R.; Sacco, A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate H-1 NMR PFG measurements. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740–4742.

(53) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(54) Groot, R. D.; Rabone, K. L. Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.* **2001**, *81*, 725–736.

(55) The value of the permittivity can vary with the condition of the simulation, size of the computational box, etc.

(56) Kusalik, P. G.; Svishchev, I. M. The spatial structure in liquid water. *Science* **1994**, *265*, 1219–1221.

(57) van Maaren, P. J.; van der Spoel, D. Molecular dynamics of water with novel shell-model potentials. *J. Phys. Chem. B* **2001**, *105*, 2618–2626.

(58) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. A systematic study of water models for molecular simulation: derivation of water models optimized for use with a reaction field. *J. Chem. Phys.* **1998**, *108*, 10220–10230.

(59) Mahoney, M. W.; Jorgensen, W. L. Diffusion constant of the TIP5P model of liquid water. *J. Chem. Phys.* **2001**, *114*, 363–366.

(60) Yu, H.; van Gunsteren, W. F. Charge-on-spring polarizable water models revisited: from water clusters to liquid water to ice. *J. Chem. Phys.* **2004**, *121*, 9549–9564.

(61) Yu, H.; Hansson, T.; van Gunsteren, W. F. Development of a simple self-consistent polarizable model for liquid water. *J. Chem. Phys.* **2003**, *118*, 221–234.

(62) Chen, F.; Smith, P. E. Simulated surface tensions of common water models. *J. Chem. Phys.* **2007**, *126*, 221101–221104.

(63) Wang, H.; Junghans, C.; Kremer, K. Comparative atomistic and coarse-grain study of water: what do we lose by coarse-graining. *Eur. Phys. J. E* **2009**, *28*, 221–229.

(64) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(65) Murrell, J. N.; Jenkins, A. D. *Properties of liquids and solutions*, 2nd ed.; John Wiley & Sons: Chichester, U. K., 1994; pp 1−299.

(66) Eisenberg, D.; Kauzmann, W. *The Structure and Properties of Water*; Oxford University Press: Oxford, U.K., 1969; pp 1−308.

(67) Dilmohamud, B. A.; Seeneevaseen, J.; Rughooputh, S. D. D. V.; Ramasami, P. Surface tension and related thermodynamic parameters of alchohols using the Traube stalagmometer. *Eur. J. Phys.* **2005**, *26*, 1079.

(68) Rodnikova, M. N. A new approach to the mechanism of solvophobic interactions. *J. Mol. Liq.* **2007**, *136*, 211–213.

(69) Kalcher, I.; Horinek, D.; Netz, R. R.; Dzubiella, J. Ion specific correlations in bulk and at biointerfaces. *J. Phys.: Condens. Matter* **2009**, *21*, 424108.

(70) Shotton, M. W.; Pope, L. H.; Forsyth, V. T.; Langan, P.; Grimm, H.; Rupprecht, A.; Denny, R. C.; Fuller, W. A high-angle neutron fiber diffraction study of the hydration of B-DNA. *Physica B* **1998**, *243*, 1166–1168.

(71) Young, M. A.; Ravishanker, G.; Beveridge, D. L. A 5-ns molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. *Biophys. J.* **1997**, *73*, 2313–2336.

(72) Cheatham, T. E., 3rd.; Srinivasan, J.; Case, D. A.; Kollman, P. A. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. *J. Biomol. Struct. Dyn.* **1998**, *16*, 265–280.

(73) Duan, Y.; Wilkosz, P.; Crowley, M.; Rosenberg, J. M. Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.* **1997**, *272*, 553–572.

(74) Feig, M.; Pettitt, B. M. Modeling high-resolution hydration patterns in correlation with DNA sequence and conformation. *J. Mol. Biol.* **1999**, *286*, 1075–1095.

(75) Young, M. A.; Beveridge, D. L. Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn. *J. Mol. Biol.* **1998**, *281*, 675–687.

(76) Kochoyan, M.; Leroy, J. L. Hydration and solution structure of nucleic acids. *Curr. Opin. Struct. Biol.* **1995**, *5*, 329–333.

(77) 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.* **1981**, *78*, 2179–2183.

(78) Shui, X.; Sines, C. C.; McFail-Isom, L.; VanDerveer, D.; Williams, L. D. Structure of the potassium form of CGC-GAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry* **1998**, *37*, 16877–16887.

(79) 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* **1998**, *37*, 8341–8355.

(80) Manning, G. S. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Q. Rev. Biophys.* **1978**, *11*, 179–246.

(81) Ponomarev, S. Y.; Thayer, K. M.; Beveridge, D. L. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14771–14775.

(82) Savelyev, A.; Papoian, G. A. Electrostatic, steric, and hydration interactions favor Na+ condensation around DNA compared with K+. *J. Am. Chem. Soc.* **2006**, *128*, 14506–14518.

(83) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(84) McConnell, K. J.; Beveridge, D. L. Molecular dynamics simulations of B-DNA: sequence effects on A-tract-induced bending and flexibility. *J. Mol. Biol.* **2001**, *314*, 23–40.

(85) Feig, M.; Pettitt, B. M. Sodium and chlorine ions as part of the DNA solvation shell. *Biophys. J.* **1999**, *77*, 1769–1781.

(86) Fuoss, R. M.; Katchalsky, A.; Lifson, S. The potential of an infinite rod-like molecule and the distribution of the counter ions. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 579–589.

(87) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. A coarse grain model for phospholipid simulations. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

CT100379F

Chapter 11

Coarse Grain Potentials: A model for DNA in Implicit and Explicit Solvent.

Pablo D. Dans,

Holds a PhD in chemistry. Staff researcher at the Institut Pasteur de Montevideo (2008-2011) andcurrently postdoctoral researcher at the Institute for Research in Biomedicine of Barcelona, Spain. Area of expertise: Theoretical biophysics.E-mail: pdans@mmb.pcb.ub.es

Leonardo Darré,

PhD student of Biophysics at the Group of Biomolecular Simulations, Institut Pasteur de Montevideo,

E-mail: ldarre@pasteur.edu.uy

Matías R. Machado,

PhD student of Biophysics at the Group of Biomolecular Simulations, Institut Pasteur de Montevideo,

E-mail: mmachado@pasteur.edu.uy

Ari Zeida,

PhD student at the Department of Chemistry, University of Buenos Aires, Argentina

azeida@qi.fcen.uba.ar,

and Sergio Pantano[*]

Institut Pasteur de Montevideo, Uruguay

* spantano@pasteur.edu.uy

Keywords: Coarse Grain, Electrostatics, Nucleic Acids, Molecular dynamics, Simulations, Solvation.

web resources:

databases:

- http://mmb.pcb.ub.es/microsecond
- (PDB, http://www.pdb.org).

software:

- (http://ambermd.org/
- (http://www.gromacs.org/).

11.1. Introduction.

Biomolecular simulations face the challenge of representing systems of enormous complexity, which are characterized by disparate size and time scales. Fully atomistic molecular dynamics (MD) techniques presented in previous chapters have become a trustworthy tool for the structural, dynamical and functional characterization of macromolecular systems. However, their applicability is restricted to relatively small systems and/or short simulation times. In fact, all-atom MD simulations on very large systems require large clusters of hundreds of computers and, in consequence, cannot easily be studied by traditional strategies. Similarly, simulations of processes on long timescales like proteins and nucleic acids folding and unfolding are prohibitively expensive, because they require so many time of calculation (months, years) and simultaneously several computer processors (CPU or GPU) reaching the limits of the actual performance of parallel computing. The reduction of systems' complexity presents an interesting option to reach longer simulation times in bigger molecular systems. Therefore, substantial effort has been devoted to the implementation of simulation techniques based on simplified or coarse grain (CG) representations of atomic systems, and thus, several CG models have been developed for investigating the longer time- and length-scale dynamics that are critical to many biological processes involving membranes, proteins, nucleic acids and the interactions between them. In these schemes, the number of atoms composing a given molecular system is drastically reduced, significantly decreasing the computational cost but keeping the physical essence of the phenomena under examination. A simpleexample of coarse graining is the united-atom approximation, which is still in use for some simulations of biological systems. In the coarse grain version of the very popular OPLS force field instead of treating all four atoms of a CH_3 methyl group explicitly one represents the whole group with a single bead (in general this contraction scheme is applied in order to avoid the explicit representation of all nonpolar hydrogen). This new bead, which has less physical meaning should, of course, be properly parameterized to represent the complete methyl group. The parameterization of these coarse grained models must be done empirically, by matching the behavior of the model to appropriate experimental data or higher detail calculation (all-atom or even quantum methods). When coarse

graining is done at higher levels (i.e. one bead per biological residue or even one bead per polymer), the accuracy of the description may be less reliable in describing some properties, and the parameterization is less and less intuitive since most of the chemistry we know from atoms in molecules are lost. No matters the granularity of the system, this kind of coarse grained potentials should be always tested thoroughly and sometimes fine-tuned to reproduce target properties at the desired level of accuracy.

In this chapter we will focus on a recently developed CG scheme for simulating DNA (see the recommended reading section). This model has the advantage of being fully back-mappable, i.e., the atomistic coordinates can be recovered from the CG trajectories with a rough root mean square deviation (RMSD) of ~2 Å, allowing for a direct comparison with atomistic data. In order to provide a comprehensive description of the model, its potentiality and limitations, it will be shown how it is possible to compare MD simulations performed at the CG level using implicit and explicit solvent with state-of the-art all-atom (AA) simulations for nucleic acids and also with experimental structures reported in the Protein Data Bank. For the comparison, we will use as main benchmark system a double helical segment in B-DNA conformation known as the Drew-Dickerson (DD) dodecamer. This dodecamer of sequence 5′-d(CGCGAATTCGCG)-3′ has been largely studied by means of experimental and theoretical works, giving rise to a solid bibliographic base to compare the CG model Having established the performance of the model, it will be used to predict new physical-chemical properties that can only be measured in the new timescale we can now reach. Particularly, the solvent and ions atmosphere formed around the DNA will be discussed as a new emerging property of this specific CG model. Futhermore, it will be shown that the CG simulations of the DD dodecamer in explicit solvent at different ionic strengths show a reversible ion condensation on the minor groove of the DNA. This ion condensation produces a conformational change induced by the concomitant interactions between cations and the minor groove of the double helix producing a sizeable narrowing in the DNA, in agreement with crystallographic data. This conformational effect is only plenty appreciable at the multi-microsecond time and could not be observed by means of atomistic simulationsscale. This provides a clear example of the usefulness of simplified approaches, which may allow for a better understanding of

structural and dynamical features of biomolecular assemblies, expanding the current limits of fully atomistic simulation techniques but keeping the essence of the chemical/physical description.

11.2. Derivation of the CG Models.

Generally speaking, the process of creating a CG representation for an arbitrary molecular system consists of the following steps:

i) **Define mapping rules between AA and CG schemes**. This implies deciding which parts of the atomistic system will be kept in the simplified representation. These can correspond to actual atoms, to a set of centers of mass or charge or to a certain groups of atoms, etc. For instance, a common choice for protein systems is to retain the coordinates of the central (Cα) carbon of each amino acid (see chapter 19 in ref. 1). Using this mapping rule, the CG representation of a protein would be a polymer of effective beads.

ii) **Define the interacting potential**. Once the mapping has been established, it is necessary to define how the effective beads interact among them. In the example of the protein coarse grained to the Cα positions, one has to decide the distances and angles between consecutive beads, their masses, sizes and eventually non-bonded or long-range interactions. All these choices are rather arbitrary and depend of the kind of phenomena of interest.

The derivation of interaction parameters can be approached using systematic or heuristic methods. Some methods incorporate data from atomistic simulations (see chapters 3, 6 and 7 in ref. 1), empirical fitting to thermodynamics data (see chapters 2, 8 and 22 in ref. 1), fitting to a experimental structural data (chapters 11, 21 and 25 in ref. 1), etc.

In our case, the CG model for DNA was derived to fit structural properties of the B-form of DNA. We use only six atoms per nucleobase corresponding to the phosphorus, the carbons at positions C5' and C1', and three atoms in the Watson-Crick region (Figure 11.1 a, b). This model was derived for simulations using implicit solvation within the generalized Born model approximation or in presence of an explicit aqueous solvent developed in our group, which includes the most biologically relevant monovalent electrolytes (Na^+, K^+, and Cl^-). CG water molecules (called Wat Four or WT4 for brevity) are represented by four linked beads (Figure 11.1 c). Each of the four beads carries a partial charge (two

positive and two negative), while single electrolytes are represented by charged spheres with a volume and mass corresponding to that of its atomistic counterparts together with its first solvation shell.

In both solvation schemes the use of the same standard form of the classical Hamiltonian allows for a straightforward implementation of our CG model in common molecular dynamics packages. A complete list of the interaction parameters is reported in Table 11.1.

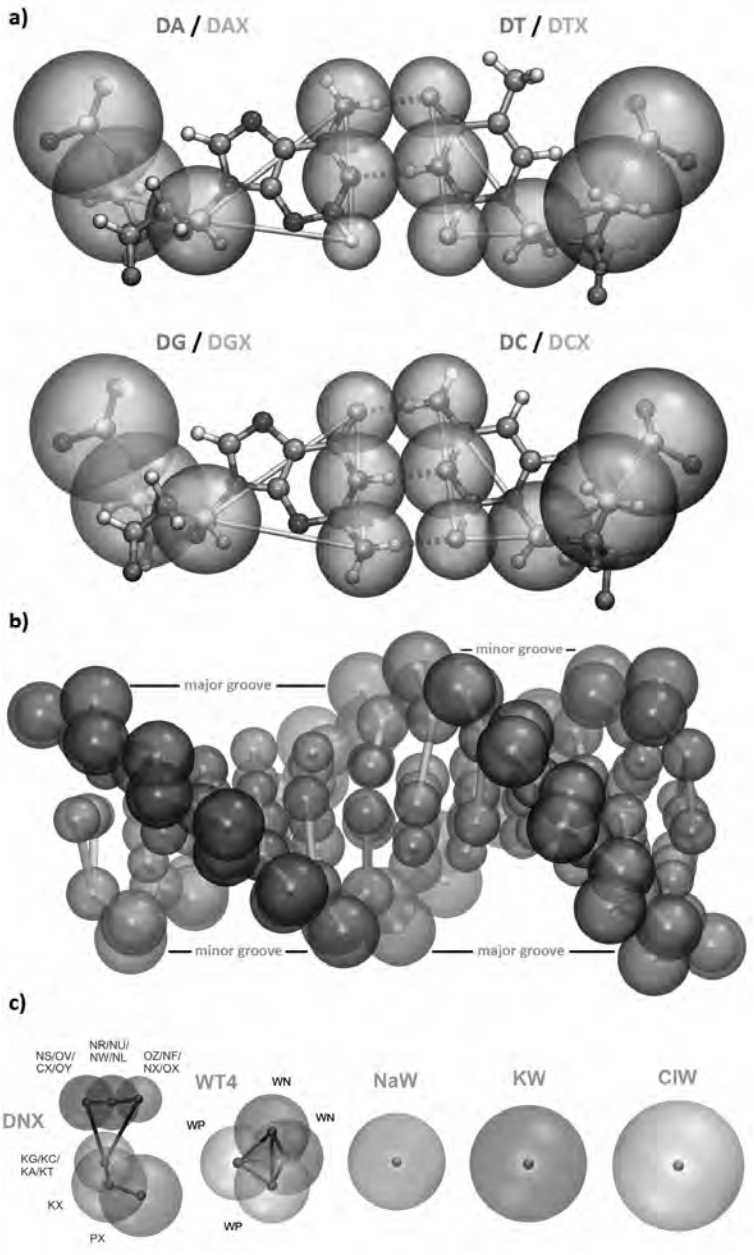

Figure 11.1. Coarse grain schemes of nucleotides and solvent molecules. The superatom types as used in the parameters and topologies files are provided in the on line tutorial. a) Adenine-Thymine and Guanine-Cytosine Watson-Crick base pairs at atomistic (solid) and CG (semi transparent) levels. The sizes of the CG beads correspond to their actual van der Waals radii. b) Side view of the Drew-Dickerson dodecamer at CG level. The 5'-3' polarity and minor/major grooves can be clearly

recognized. c) Comparative view of all the molecules in the CG scheme. DNX stands for DGX, DCX, DAX or DTX, the four CG nucleobases. The semi transparent spheres provide an idea of van der Waals radii on the same relative scale between the CG particles.

2.1 Computational Details.

The functional form of the Hamiltonian used is:

$$V = \sum_{\text{bonds}} \mathrm{k_b}\left(r_{ij}\text{-}\mathrm{r_{eq}}\right)^2 + \sum_{\text{angles}} \mathrm{k_\theta}\left(\theta\text{-}\theta_{\mathrm{eq}}\right)^2 + \sum_{dihedrals} \frac{\mathrm{V_k}}{2}\left[1+\cos\left(\mathrm{n_k}\varphi\text{-}\gamma_{\mathrm{k}}^{\mathrm{eq}}\right)\right]$$
$$+\sum_{\mathrm{l}}^{\mathrm{N}}\sum_{\mathrm{l>m}}^{\mathrm{N}}\left\{\varepsilon_{\mathrm{lm}}\left[\left(\frac{\sigma_{\mathrm{lm}}}{r_{lm}}\right)^{12}-\left(\frac{\sigma_{\mathrm{lm}}}{r_{lm}}\right)^{6}\right]+\frac{\mathrm{q_l q_m}}{4\pi\varepsilon_0\varepsilon_{\mathrm{r}} r_{lm}}\right\} \quad (1)$$

where k_b is the bond stretching constant, $r_{ij} = r_i - r_j$, and r_{eq} is the equilibrium bond distance between two linked elements. k_θ is the bond angle constant, θ is the instantaneous angle defined by three successive elements and θ_{eq} is the equilibrium bond angle. V_k is the height of the torsional barrier, n_k is its periodicity, φ is the torsion angle defined by four consecutively bonded elements, γ_k^{eq} is the phase angle, and N is the number of particles of the system. The last term corresponds to the Lennard-Jones and Coulombic potentials, in which ε_{lm} is the maximum depth of the function, σ_{lm} is the distance from the atomic center to the minimum of the van der Waals function, q is the net point charge of each bead, ε_0 is the vacuum permittivity, ε_r the dielectric constant, and r_{lm} is the distance between two beads.

Table 1. Set of interaction parameters for the CG models.

Mass, charge and Lennard-Jones parameters for DNA and WT4												
Superatom type[a]	Mass	Charge q (e)		Lennard-Jones ε_{lm} (Kcal/mol)	σ_{lm} [b] (Å)							
PX	51.53	-1.00		3.5035	4.6327							
KX	51.53	0.00		1.9164	4.2906							
KA	51.53	0.00		1.9164	3.3997							
NX	51.53	0.20		4.2042	3.2500							
NW	51.53	-0.20		4.2042	3.2500							
CX	51.53	0.00		1.9164	2.6699							
KT	51.53	0.00		1.9164	3.3997							
OX	51.53	-0.20		4.2042	2.9599							
NL	51.53	0.40		3.3283	3.2500							
OY	51.53	-0.20		4.2042	2.9599							
KG	51.53	0.00		1.9164	3.3997							
OZ	51.53	-0.40		5.4304	2.9599							
NR	51.53	0.20		4.5546	3.2500							
NS	51.53	0.20		4.5546	3.2500							
KC	51.53	0.00		1.9164	3.3997							
NF	51.53	0.40		4.5546	3.2500							
NU	51.53	-0.20		4.5546	3.2500							
OV	51.53	-0.20		5.4304	2.9599							
WN	50.00	-0.41		0.1314	4.2000							
WP	50.00	0.41		0.1314	4.2000							
NaW	130.99	1.00		0.1314	5.8000							
KW	147.10	1.00		0.1314	6.4500							
ClW	143.45	-1.00		0.1314	6.8000							
Bond and angle force constants for DNA												
k_b [c]	50 Kcal/mol•Å^2			k_θ [d]	75 Kcal/mol•rad^2							
Bond parameters for WT4												
k_b	5 Kcal/mol•Å^2			r_{eq}	4.5 Å							
Dihedral parameters for DNA												
	V_1 [f]	V_2	V_3	V_4	n_1	n_2	n_3	n_4	γ_1^{eq} [g]	γ_2^{eq}	γ_3^{eq}	γ_4^{eq}
KN[e]-PX-KX-KN(Φ)	3.75				8				161.0			
PX-KX-KN-PX (Ξ)	3.75				8				-153.2			
KX-KN-PX-KX (Ψ)	3.75				4				-29.3			
PX-KX-KA-NX (Ω_{DAX})	3.75	2.25	2.63	3.75	1	7	2	1	118.0	47.0	20.0	-220.0
PX-KX-KA-CX (Γ_{DAX})	2.25	1.50	0.75		1	3	4		65.0	145.0	130.0	
PX-KX-KT-OX (Ω_{DTX})	3.75	1.88	2.63	3.75	1	8	2	1	117.0	47.0	20.0	-140.0
PX-KX-KT-OY (Γ_{DTX})	2.25	1.50	0.75		1	3	4		65.0	145.0	130.0	
PX-KX-KG-OZ (Ω_{DGX})	3.75	2.44	2.63	3.75	1	6	2	1	110.0	90.0	20.0	-220.0
PX-KX-KG-NS (Γ_{DGX})	2.25	1.50	0.75		1	3	4		65.0	145.0	130.0	
PX-KX-KC-NF (Ω_{DCX})	3.75	1.88	2.63	3.75	1	8	2	1	117.0	47.0	20.0	-140.0
PX-KX-KC-OV (Γ_{DCX})	2.25	1.50	0.75		1	3	4		65.0	135.0	130.0	

[a] The type of the superatoms match those included in the parameters and topology files available in the book's web site. [b] Distance from the atomic center to the minimum of the vdW function. [c,d] All the equilibrium distances (r_{eq}) and angles (θ_{eq}) were set in correspondence with the inter-bead distances and angles measured from the canonical B-DNA. [e] Where KN = KA, KT, KC or KG. [f] Rotational barriers in Kcal/mol. [g] Phase angles in degrees.

2.1.1 Simulation protocols.

Along this paragraph we discuss the most salient technical details to set up and run a simulation. Esentially, the steps needed to run calculations using the simplified CG model presented here are the same used for any MD simulation. In the example shown in the tutorial, MD simulations are performed in implicit solvation using the generalized Born (GB) approximation as implemented in the AMBER10 suite of programs. In general, for short DNA fragments it may be convenient to include restraints to the capping base pairs as during long simulations the double helical structure can partially separate at the ends (see on line tutorial for technical details).

Besides the implicit solvation approach, the same set of parameters can be used to perform simulations embedding the solute in an explicit CG solvent called WatFour (WT4, for shortness). These explicit solvent simulations are not described in the on line tutorial and were performed using GROMACS. Input parameters and technical details can be found in Darre' et al 2010 (see reference section).

2.1.2 Simulated systems.

For systems 1 and 2 (sys1 and sys2, Table 2), the crystal structure of the Drew-Dickerson dodecamer with PDB code 1BNA was used as starting structure. One can then compare the results with state-of-the-art all-atom simulations(hereafter called sys3).. Sys1 and sys2 were first run for 1.2 μs, and snapshots were collected for analysis every 10 ps.

To analyze the dynamical behavior of the Drew-Dickerson DNA embedded in a different ionic strength, the simulation of the sys2 was then extended, sampling every 50 ps, for a total simulation time of 12 μs. Additionally, we also extended up 12 μs the simulation presented in Darré *et al* (sys4).[3] Finally, two sequences were chosen to study the accuracy of the model to reproduce the intra-base pair helical parameters of the 10 unique dinucleotide steps: AA·TT, AC·GT, AG·CT, AT·AT, CA·TG, CC·GG, CG·CG, GA·TC, GC·GC and TA·TA. The duplexes, were constructed in the canonical B-DNA form: GCCTATAAACGCCTATAA (sys5), and CTAGGTGGATGACTCATT (sys6). Both systems were run for 10 μs in explicit solvent using an ionic concentration of roughly 0.2 M. The results were

compared with two recent state-of-the-art all-atom force fields simulations: parm99-bsc0 and charmm27 (for details see the related information at the web sites http://ambermd.org and http://www.charmm.org/).

2.1.3 Tools for analysis.

Coordinates' back-mapping and analysis where performed using *in-house* programs (provided in the on line tutorial) and free software (Curves+ (http://gbio-pbil.ibcp.fr/Curves_plus/Curves+.html), PTRAJ module of Ambertools 1.2, and GROMACS utilities).

Convergence and stability of all the studied systems were evaluated by means of Root Mean Square Deviations (RMSD), B-factors, major and minor groove dimensions, and dipole moments. The essential dynamics of the different duplexes were derived by diagonalization of the covariance matrix. The essential modes were extracted for all the heavy atoms.

Table 2. Description of the simulated systems.

System	Solvation model	nº solvent molecules [a]	Ionic Species (nº of ions)	Nucleotide sequence (5'-> 3')	time (μs)	[Ion Conc.] (M)
Sys1	GB	---	---	CGCGAATTCGCG [b]	1.2	0.15 [c]
Sys2	WT4	523 (5753)	NaW+(22)	CGCGAATTCGCG	1.2 / 12	0.21
Sys3[d]	TIP3P	4998	Na+(22)	CGCGAATTCGCG	1.2	0.24
Sys4	WT4	506 (5566)	NaW+(19) KW+(19) ClW-(16)	CGCGAATTCGCG	12	0.36
Sys5	WT4	1510 (16610)	NaW+(34) KW+(33) ClW-(33)	GCCTATAAACGCCTATAA	10	0.21
Sys6	WT4	1510 (16610)	NaW+(34) KW+(33) ClW-(33)	CTAGGTGGATGACTCATT	10	0.21

[a] Parentheses indicate the equivalent number of AA water molecules represented [b] Drew-Dickerson Dodecamer.
[c] Considered through the linearized Debye-Hückel approximation. [d] Taken from http://www.mmb.pcb.ub.es/microsecond.

To access the similarity in essential movements between two simulations, trajectories were fitted to a common reference (the canonical structure) and compared using the similarity index (SI):

$$SI = 1 - \sqrt{\left[tr\left(\sqrt{\frac{M1}{tr(M1)}} - \sqrt{\frac{M2}{tr(M2)}} \right) \right]^2} \qquad (2)$$

Where *M1* and *M2* are two covariance matrixes. The SI is 1 for identical matrices and 0 when the sampled subspaces are orthogonal. Essential dynamics analysis was done with the G_COVAR and G_ANAEIG modules of GROMACS 4.0.5 and also with the PCASUITE program.

The cation-induced narrowing of the minor groove was studied for the DD sequence. Such structural changes were estimated from the average inter-phosphate distance between opposite strands (including the capping base pairs) measured for the following pairs: {(*5,* 24), (*6*, 23), (*7*, 22), (*8*, 21), (9, *20*), (10, *19*), (11, *18*), (12, *17*)} (italics indicate the residue numbers at the AT track). Cations were considered to be bound to the minor groove if their distance to the phosphate groups of both opposite strands was below 5 Å. When the amount of cations was measured for the central track, only the four central phosphate pairs were considered. The major groove size was estimated from the average inter-phosphate distance between opposite strands measured for the following pairs: {(2, 19), (3, 18), (4, 17), (5, 16), (6, 15), (7, 14)}. Finally, WT4 beads were considered bound to the DNA backbone when the WP-PX distance was below 5.5 Å.

For the calculation of several properties the CG trajectories were back-mapped to recover the atomic coordinates. This permitted to evaluate the overall structural quality of the DNA dodecamer in terms of helical parameters, compare directly the computed B-factors with those obtained in X-ray experiments and compare the essential dynamics with state-of-the-art all-atom simulations.

3 Advantages, Limitations and Perspectives of the CG model for DNA.

In the following paragraphs we discuss the use of the CG model for DNA simulation organized in three main topics: i) Structural and dynamical comparison of the CG-DNA model in implicit and explicit solvents against atomistic simulations. ii) The sequence specific base pair step conformations as obtained from explicit solvent CG simulations of double stranded DNA containing the 10 unique base pairs were compared against AA simulations performed with popular force-fields for nucleic acids, averaged experimental results, and canonical A/B-DNA. iii) Finally, we used the explicit solvation approach to study the electrostatics-driven effects of the solvent atmosphere on the double helix. Significant sequence dependent bending events occurred due to specific and reversible DNA – electrolyte binding in multi-microsecond timescale.

3.1 Structural and dynamical comparison.

3.1.1 Recovery of atomistic information from the CG trajectories.

Aimed to establish the capabilities and limitations of both solvation approaches when using the CG model we first present a systematic comparison of simulations of the DD dodecamer performed with implicit and explicit schemes (sys1 and sys2, respectively). To allow a proper comparison with AA simulations reported by Pérez *et al* [4] (sys3) and experimental structures (PDB codes 1BNA and 1FQ2), the entire trajectories were fully back-mapped.

It is worth to note that the back-mapping procedure tends to homogenize certain sub-states observed during the all-atom simulations. After back-mapping, the ζ/ε torsions are always in the BI conformer and α/γ torsions are re-constructed in the canonical g-/g+ distribution. Experimentally, near 15% of the ζ/ε torsions are in the BII conformer, and some less frequent sequence dependent shifts for the α/γ torsions are also observed. In out case, the sugar pucker is always reconstructed in the C2'-endo typical of the canonical B-form. Nevertheless, as a result of the energy optimization performed as final step of

the back-mapping procedure, we obtain 80% of C2'-endo conformations, while the remaining conformers correspond mainly to C3'-exo.

3.1.2 Comparison between CG and atomistic simulations.

Both solvation schemes described equally well the DD structure with no RMSD drift from the experimental structure (PDB code:1BNA) along the 1,2 μs explored (Figure 2a). The Drew-Dickerson DNA duplex showed in all simulations a stable but very flexible structure oscillating around the equilibrium B-form. The higher number of conformational substates explored by the AA approach translates in a higher RMSD. From a dynamic point of view, both CG approaches furnished a similar picture of the structural behavior of the double helical segment of DNA with a maximum pair wise RMSD between both averaged structures of 2.5 Å. This maximum separation was observed at a simulation time around 0.6 μs (Figure 2a). These peaks, as well as those near 0.75 μs, 0.8 μs and 0.9 μs, arose from structural perturbations related with the transient binding of Na+ present in the solution (see below). The observed differences were rather small as deduced from average RMSD values from AA (1.9 Å) and any of the CG simulations (1.5 Å). This can be also observed from structural superposition of snapshots taken after thermalization, at the middle and end of the dynamics (Figure 2b).

The mobility of the DNA heavy-atoms can be computed and expressed in terms of B-factors. This information can be directly compared with the AA simulation, and with values coming from X-ray experiments. Although the temperatures at which were crystallized the structures 1BNA and 1FQ2 are significantly different (290 K and 100 K, respectively), both have the same qualitative behavior. B-factors showed a more flexible pattern towards the end of each strand respect to the central part of the duplex (Figure 2c). From the profile of the peaks, it can be seen how all the simulation schemes reproduce correctly the higher mobility of the phosphate groups in each nucleobase. Each residue showed a second shoulder that corresponds to the sugar moiety, and finally the bases, which are always the less mobile part of the molecule. In absolute terms, the all-atom MD simulation exhibited the most

flexible behavior. Both, implicit and explicit solvation schemes for simplified DNA give the same quantitative results, which are intermediate between the two experimental values

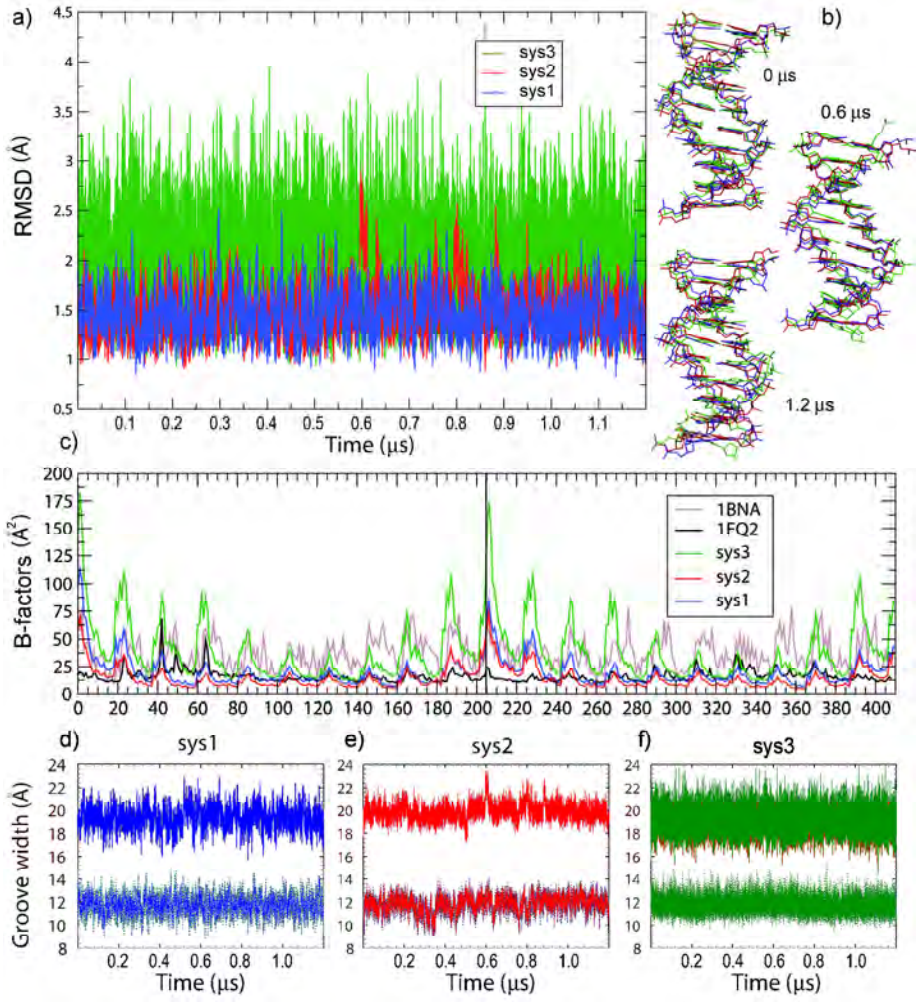

Figure 2. Comparison between AA and CG simulations. Implicit solvation CG (sys1), explicit solvation CG (sys2) and AA simulations (sys3) are presented in blue, red and green, respectively. a) RMSD evolution along time calculated for all the heavy atoms respect to the X-ray structure 1BNA. b) Least mean square fit performed on all heavy atoms of conformers taken at the beginning, middle, and end of the trajectories. c) B-factors calculated from MD simulations. The corresponding quantities from the starting X-ray structure (pink) and from the high resolution structure 1FQ2 (black) are included for comparison. The x-axis corresponds to the sequential number of the heavy atoms in both strands. The vertical line indicates the separation between both strands. The values reported for the MD simulations were calculated as $B=8/3\ \pi^2\ RMSF^2$. The RMSF were obtained over the 1.2 μs trajectory after back-mapping in sys1 and sys2 and directly from the instantaneous position in sys3. d - f) Major and minor groove widths for the three studied systems are represented by the straight and pointed lines, respectively.

The agreement between the different simulation schemes can be also evaluated from the dimensions of the minor and major grooves. We found a good agreement in all the cases (Figure 2d-f), with lower fluctuations for the CG schemes. Moreover, looking at the groove's variation in Figures 2d-f, it seems that the presence of the explicit CG solvent reduces, particularly for the major groove, the thermal oscillations. Sys2 presents transiently slightly narrower minor groove as compared with sys1 and sys3. As detailed in the next section, this is the result of a specific phosphate cation interaction inside the minor groove. The local screening of the negatively charged phosphate (PX) beads by the sodium (NaW^+) ions in sys2 results more effective than the homogeneous screening effect produced by the implicit solvent model used in sys1. This confers to sys2 the capability to sample regions of the potential energy surface, which are very rarely visited by sys1 as suggested by the number and time extension of the RMSD peaks shown in Figure 2a.

We now turn our attention to the study of the dynamical determinants of each system. We compared the conformational subspace sampled by inspecting the essential dynamics modes, which are characteristic of each simulation. To capture the 95% of the variance during the simulations 33 eigenvectors were needed for sys1, 35 for sys2 and 44 for sys3. In all the simulations, the first 3 eigenvectors explained approximately 50% of the total variance. These three essential modes were analyzed in more detail in terms of their projection onto the real space. A marked similarity between the first two modes enclosing in average 43% of the total variance was observed respect to the AA simulation. In all the cases, the first mode involved a twisting and untwisting movement correlated with changes in the grooves; the second concerned bending and twisting around the center of the AT track; and the third, represented a global tilting of the duplex.

To achieve a more quantitative characterization, a similarity index (SI) between the covariance matrices calculated along the MD trajectories. To set a reference level for the comparison, we first calculated the SI according to equation 2 (section **2.1.4**) between both halves of the fully atomistic trajectory of 1,2 μs[4], i.e., the SI was considered between covariance matrixes calculated along the first

and last 0.6 μs. This gave a value of 0.91, which might be considered as the maximum figure we can expect from any SI between CG and AA simulations.

The CG simulations performed using implicit and explicit solvation resulted in a similarity of 0.76. This strongly suggests that both approaches sample similar conformational spaces.

When compared to AA scheme, the similarity index of sys3 with the CG DNA in both solvents was 0.58 for sys1 and 0.66 for sys2. This rough identity between the essential dynamics described by AA and CG simulations allows postulating a good degree of the similarity of the potential energy surfaces sampled by both approaches. It also underlines the higher compatibility of the explicit solvent approach using the WT4 model with the AA benchmark, pointing out the importance of the explicit aqueous solvent in the dynamics and structural conformations of DNA.

3.2 Base pair steps and sequence specificity.

To evaluate the goodness of the model to reproduce sequence-dependent structural modifications we simulated systems 5 and 6 using explicit solvation, which contain all the unique dinucleotide base pair steps. The helical parameters were compared upon back-mapping with canonical, averaged experimental values, and results coming from AA force fields for nucleic acids simulation (parm99-bsc0 and charmm27, Figure 3). This is important to establish the capability of the model to reproduce sequence specific structural patterns. [5]

In general, the agreement with experiment and state-of-the-art AA force fields was very good although a slight flattening of the values was observed. These relatively small differences can be ascribed to the reduced set of coordinates that leads to fewer degrees of freedom and the substates homogenization, especially of the sugar pucker and backbone which is intrinsic to the back-mapping procedure.

Only the roll showed little deviation from the atomistic results (Figure 3). This is not unexpected as our model was fitted to reproduce the canonical B-DNA, which has a slightly negative roll value. Conversely, AA force fields sampled more frequently positive roll values, reproducing with less forcefulness the averaged X-ray values for the GC, AT, CA and TA steps, which exhibit negative roll

angles. Since changes in the roll parameter are involved with bending events, which are in turn crucial to study DNA – protein interactions, further application using this model should take into account these structural aspects.

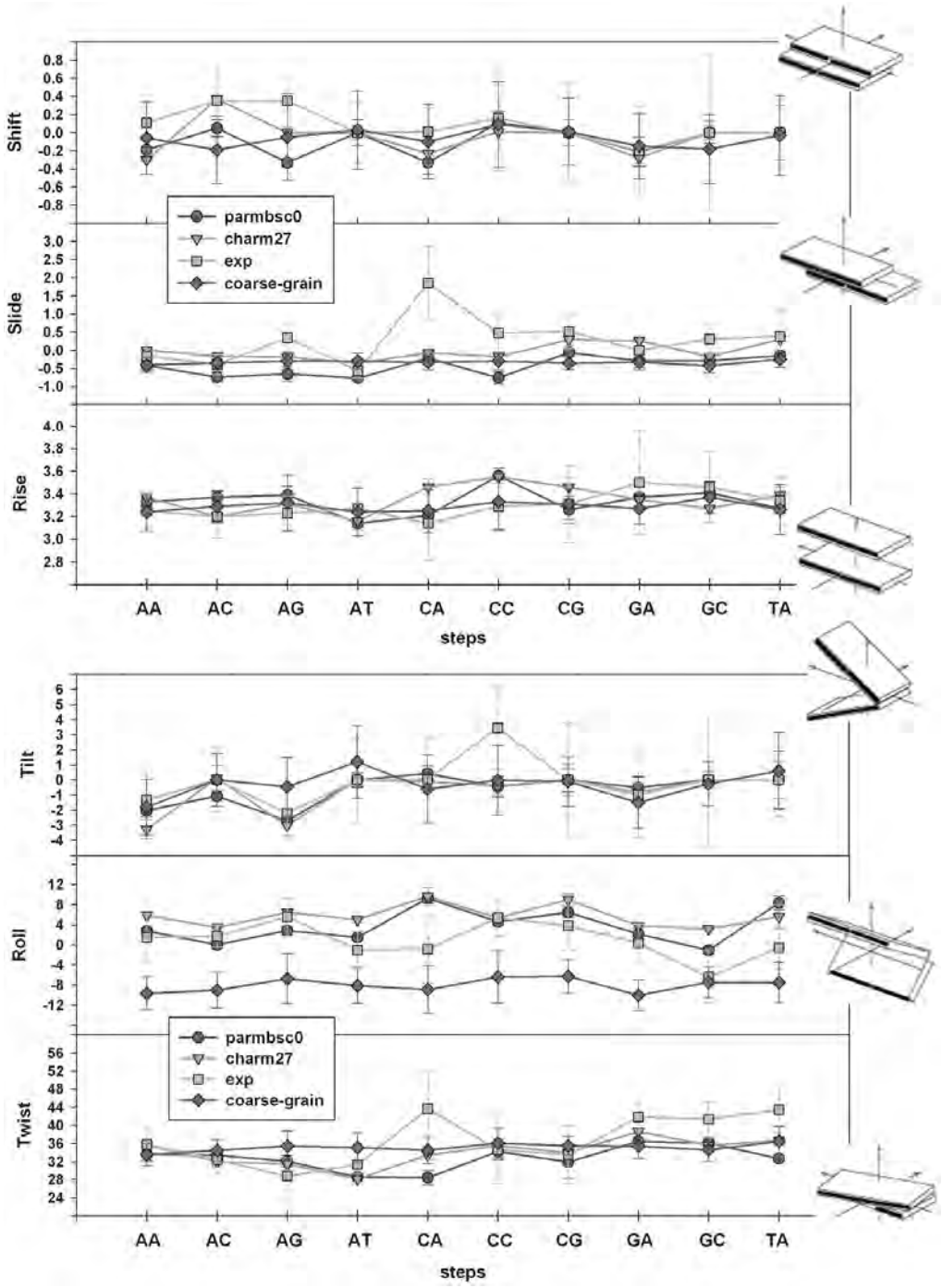

Figure 3. Averages and standard deviations of helical parameters for the 10 different dinucleotide steps. Translational parameters (Shift, Slide and Rise) shown on the top are measured in Å, while rotational ones (Tilt, Roll and Twist) are reported in degrees in the bottom panel. The parm-bsc0 (circles), charmm27 (triangles), X-ray (squares), and

results coming from the CG models (rhombi) are shown. Complementary steps have the same average except for a change in sign of shift and tilt. The parmbsc0 and charm27 values were taken from ref. 5.

3.3 Electrostatics and solvent atmosphere.

3.3.1 Solvation structure around DNA.

The solvation structure around DNA molecules is very well characterized. In particular, cones of hydration around the phosphate groups, spines of hydration inside the minor groove and an extended hydration of the major groove have been characterized by a number of early studies. Although a quantitative comparison between atomistic and CG solvents is not possible, we noticed a close similarity between the solvation structure around the double helical filaments in both cases.[3] As an example of this, we compare the arrangement of water and WT4 around the phosphate moieties. The AA phosphate groups are often surrounded by 4 to 6 water molecules clustered in two conical arrangements where each of the OP1 and OP2 oxygen atoms occupy the tip of an imaginary the cone and 2 or 3 water oxygen are at the plane of the base forming hydrogen bonds with the phosphate's oxygen (Figure 4a). At the CG level, WT4 particles surround the PX superatoms generating a nearly conical spatial arrangement (Figure 4b). Nearly 4 positive WT4 beads (Figure 4c) are surrounding the CG phosphate moiety, suggesting that the electrostatics-driven solvent atmosphere produced by the WT4 particles is similar to the atomistic one.

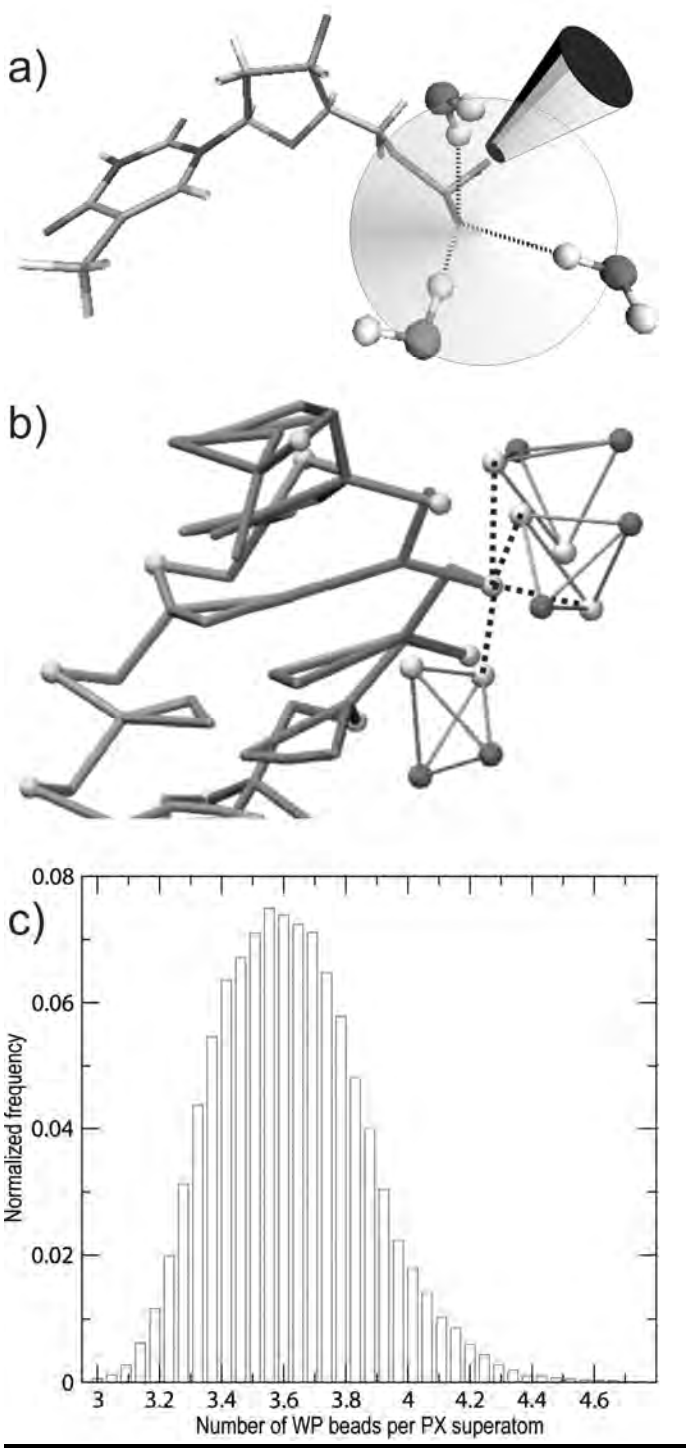

Figure 4. Solvent atmosphere around the DNA. a) Schematic representation of the cones of hydration formed around the phosphate groups in atomic solvent. b) Same as a for CG simulation. The picture shows a representative snapshot taken from the simulation of sys2. c) Number of positive WT4 beads (WP) interacting with each PX superatom, averaged over the all macromolecule and the entire simulation.

3.3.2 Electrostatics of the CG model.

The solvent organization is driven by the direct charge-charge interaction between the phosphate moiety and solvent molecules. However, higher order electrostatic interactions may also contribute to modulate the solvation structure and DNA conformation. In fact, we found a qualitatively good correspondence between the dipole moments of each individual nucleobase at the CG level when compared with the analogous quantities calculated using parm99-bsc0 or charmm27 force fields (Figure 5). The orientations of the individual dipole moments of each of the CG nucleobases are nearly collinear with the AA ones although the modules of the vectors have a larger value (Figure 5a,b). This quasiquantitative agreement is particularly relevant if we remind that dispersion forces are crucial in keeping the base-base stacking. The dipolar component resulting from the charge distribution in our model contributes to generate sequence-specific effects (Figure 3) without imposing *ad hoc* base-base or base pair-base pair potentials.

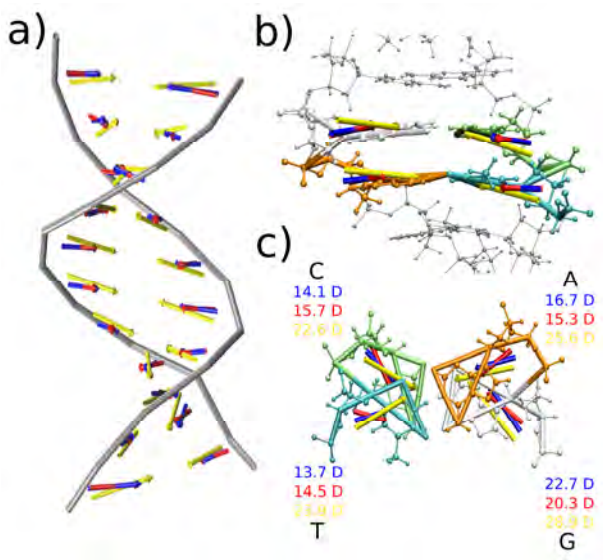

Figure 5. Dipole moments. a) Schematic representation of the backbone for the X-ray structure 1BNA, with parm99-bsc0 dipoles (blue arrows), charmm27 dipoles (red arrows), and the dipoles from the CG nucleobases (yellow arrows). b) Front view of the double helix highlighting the dipole moments for AT and GC base pairs. c) Same as b) but viewed from the bottom indicating the dipoles of each type of nucleobase in Debyes. Values in blue, red and yellow correspond to parm99-bsc0, charmm27 and CG model, respectively.

3.3.3 Electrolyte binding and DNA conformation.

In this section we analyze multi microseconds long simulations of the DD dodecamer and it is interactions with the aqueous environment. Before going in specific details, it is important to mention that up to a couple of capping base pairs might experience a rupture of the Watson-Crick interactions already within the sub microsecond range. Therefore, it cannot be excluded that for longer simulation times than those explored here, a complete melting of the relatively small double helix could eventually be observed. This effect can be avoided by applying distance restraints to the capping bases. The simulations presented hereafter were performed without any restraints.

While the global distribution of cations around the DNA contributes to the stability of the double helix, the specific interaction of cations with DNA is related with local structural distortions. The binding of one single ion may be enough to induce a sensible change in the minor groove.[3] Upon successive incorporation of ions, the narrowing becomes more marked reaching a minimum when up to seven ions are concomitantly bound. X-ray structures of the same dodecamer also reveal high occupancy of cations in the minor groove leading to its narrowing (see PDB structures: 355D and 428D). However, from the AA simulation the correlation between the entrance of ions in the minor groove and the narrowing effects is not totally clear. In the AA simulation performed by Pérez el al. (sys3)[4] the simultaneous occupancy of the minor groove by several ions is very uncommon, but the presence of one Na^+ is not so rare. Although the occupancy of those few cations in the minor groove has residence times of 10 to 15 ns, the global deformation produced by the cations in terms of the minor groove narrowing are very subtle and, apparently, not sufficient to explain the distortions observed in the X-ray structures. From the analysis of the corresponding CG trajectory (sys2) performed on a tenfold longer timescale, the minor groove narrowing appears to take place on two different time scales. The first is related with the binding of one or two ions for up to few dozens of nanoseconds, while the second corresponds to the simultaneous binding of several ions for microsecond long periods (Figure 6). The nanosecond binding events are in very good agreement with the residence times computed from atomistic simulations and also respect to the location of the cations in the minor groove exhibiting the

same high affinity binding site for Na^+ detected in AA simulations. Moreover these binding sites correspond also with the geometry reported for the binding sites of water, sodium and potassium within the minor groove of X-ray structures. This seems to point to the limited sampling reached by AA simulations as the main responsible for the discrepancies with the experimental data.

Aimed to get deeper insights onto this phenomenon we performed a simulation of the same DNA double helix in presence of higher ionic concentration and in presence of Na^+ and K^+ (sys4, see Table 2). Furthermore, to study a different sequence context and length, we calculated the narrowing also for sys5 and sys6 (Figure 6 b to d). It can be observed that besides the duration and frequency of the events, which seemed to be rather arbitrary, all the systems display an analogous behavior. Taking all these results as a whole, we can conclude that micro seconds long narrowing events spontaneously appear independently of the sequence or length of the DNA segment. The binding of up to 7 ions in the minor groove produce a drastic electrostatic collapse that permits both opposite strands to get closer and to generate a sensible narrowing. The discrimination of the total number of ions bound to the whole minor groove respect to those bound only within the central track (bottom panels in Figure 6 a to d) showed that one or two bound ions were not sufficient to produce long lasting narrowing. Notably, the average narrowing of 9.6 Å measured during the microsecond long events where 3 or more ions were bound coincides precisely with X-ray determinations. This suggests that an extended counterion condensation is needed to generate a significant and sustained bending of the DNA.

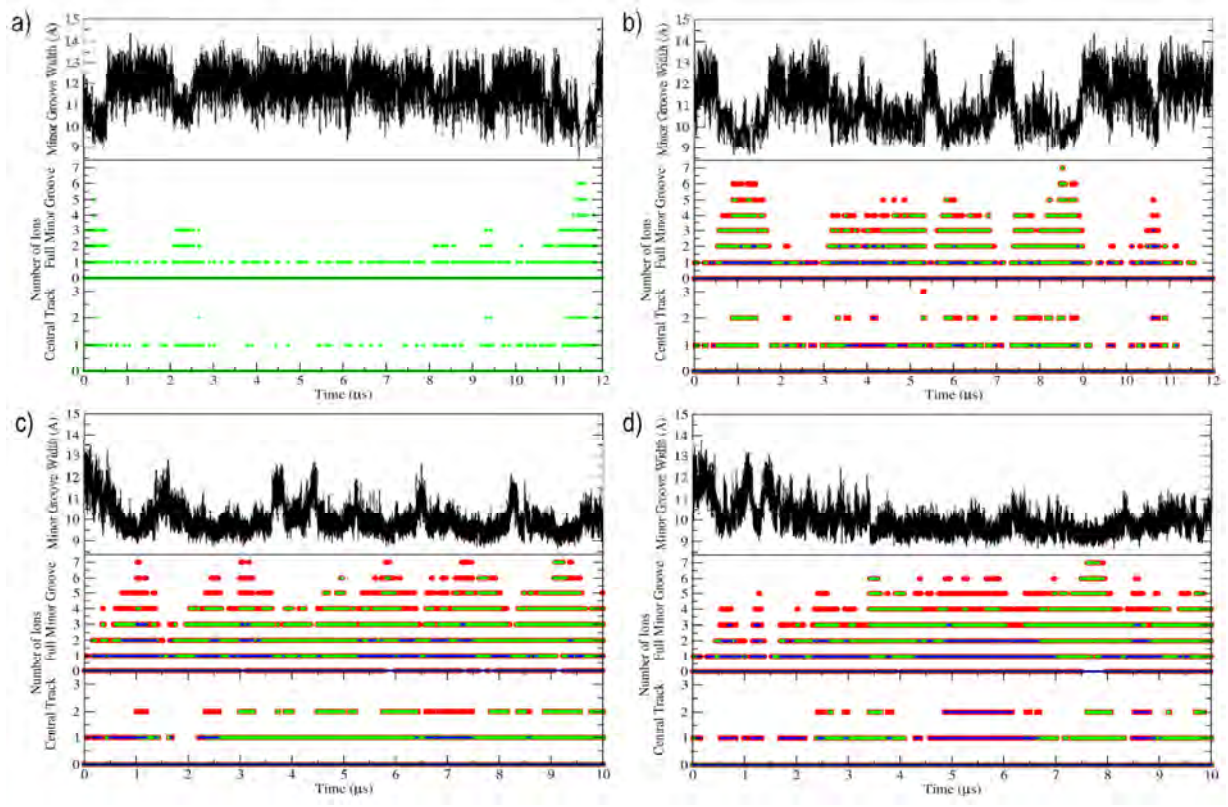

Figure 6. Binding of cations within the minor groove. a) For sys2: minor groove width (top), total number of bound cations in the minor groove (middle), and of cations bound only to the minor groove of the central track plotted against time (bottom). b) Idem than (a) for system 4. c) Idem than (a) for system 5. d) Idem than (a) for system 6. The number of cations is shown as the number of NaW^+ (green), KW^+ (blue), and the sum of both (red).

This conformational transition, corresponded to a relatively stable bending ranging from 10 to 50 degrees with and average value of 26 degrees. To put into perspective these results, we analyzed the DNA bending of all the protein-DNA complexes solved at a resolution higher than 2.5Å, as reported in the PDIdb database (http://melolab.org/pdidb/web/content/home). It comes out that 62% of these complexes display a bending lower or equal to 26 degrees, while this number increases up to 81% if we consider a bending of 50 degrees. This leaves room to speculate that protein-DNA recognition might exploit spontaneous fluctuations driven by the electrolytic environment. Alternatively, one might think that the concomitant binding of ions can create 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 comprehensive comparison of the performance of our CG model for DNA in implicit and explicit solvation against AA simulations and experimental data. It turns out that both approximations provide a reasonably good description of the structural and dynamical features of DNA. The gross determinants of the structural stability of the double helix (RMSD, B-factors major/minor groove dimensions) suggest that, upon back-mapping, the information obtained from the CG simulations is almost as accurate as that reachable with AA techniques. From the dynamical point of view, a roughly good superposition was found between the spaces sampled by both AA and CG schemes. Not unexpectedly, explicit solvent simulations using the WT4 model for CG solvation improve the agreement with AA simulations.

The good description of our CG scheme is also supported by the helical parameters resulting from the simulation of 18-mer duplexes containing the ten unique base pair steps. Comparison of these data with AA simulations and crystallographic values for the helical parameters resulted in a good agreement (Figure 3).

As a general rule, we can conclude that both implicit and explicit solvation approaches provide similar results. For large nucleic acid systems, the description provided by the implicit solvation approach is sufficiently good for nearly every structural and dynamical aspect of these macromolecules. However, solvent mediated effects (as for instance, ionic strength screening) seem to be poorly reproduced within the implicit solvation framework. While this may not pose a problem for the study of long segments of DNA, the reversible collapse of counterions and electrolyte mediated narrowing of the minor groove observed in the multi-microsecond timescale obviously requires the use of explicit solvent.

The specific and reversible binding of counterions within the minor groove generates a bend of up to 50 degrees in the double helix. 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 and the comparison with the known universe of protein-DNA complexes reported in the PDB provides a validation for our model and opens a number of

interesting questions such as: can the electrolyte environment play a role in protein-DNA recognition? Is there a sequence preference in the electrolyte condensation? May ion condensation modify binding free energy profiles?

Answering these and other questions would probably require the use of atomistic or multi scale representations. However, this is an example of how simplified methods can be useful to prospect systems' properties helping to pinpoint phenomena that because of their characteristic time lengths or sizes escape to the reach of higher-level calculations.

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).

References

(1) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, 1st ed.; Taylor & Francis Group: New York, 2009

(2) Dans, P. D.; Zeida, A.; Machado, M. R.; Pantano, S. *J. Chem. Theo. Comp.* **2010,** *6*, 1711.

(3) Darré, L.; Machado, M. R.; Dans, P. D.; Herrera, F. E.; Pantano, S. *J. Chem. Theo. Comp.* **2010**, *6,* 3793

(4) Pérez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the Microsecond Time Scale. *J. Am. Chem. Soc.* **2007,** *129*, 14739.

(5) Perez, A.; Lankas, F.; Luque, F. J.; Orozco, M. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **2008,** *36*, 2379.

PCCP

Cite this: *Phys. Chem. Chem. Phys.*, 2011, **13**, 18134–18144

www.rsc.org/pccp

PAPER

A hybrid all-atom/coarse grain model for multiscale simulations of DNA†

Matías Rodrigo Machado, Pablo Daniel Dans and Sergio Pantano*

Received 19th April 2011, Accepted 30th August 2011
DOI: 10.1039/c1cp21248f

Hybrid simulations of molecular systems, which combine all-atom (AA) with simplified (or coarse grain, CG) representations, propose an advantageous alternative to gain atomistic details on relevant regions while getting profit from the speedup of treating a bigger part of the system at the CG level. Here we present a reduced set of parameters derived to treat a hybrid interface in DNA simulations. Our method allows us to forthrightly link a state-of-the-art force field for AA simulations of DNA with a CG representation developed by our group. We show that no modification is needed for any of the existing residues (neither AA nor CG). Only the bonding parameters at the hybrid interface are enough to produce a smooth transition of electrostatic, mechanic and dynamic features in different AA/CG systems, which are studied by molecular dynamics simulations using an implicit solvent. The simplicity of the approach potentially permits us to study the effect of mutations/modifications as well as DNA binding molecules at the atomistic level within a significantly larger DNA scaffold considered at the CG level. Since all the interactions are computed within the same classical Hamiltonian, the extension to a quantum/classical/coarse-grain multilayer approach using QM/MM modules implemented in widely used simulation packages is straightforward.

Introduction

The constant struggle to perform molecular simulations of bigger and more intricate biological systems at relevant size and time scales has prompted the development of several models involving different levels of complexity to reduce the computational cost. In this sense, simplified or coarse grain (CG) representations have emerged as a powerful strategy to extend time and size scalability on complex molecular systems.[1] The idea behind any CG model is to describe the behavior of the system retaining the most relevant molecular features and interactions.[2] This generally implies reducing degrees of freedom of a given molecular system by condensing atomic information into a reduced number of effective interaction points. By accepting the loss of all-atom (AA) details these simplified models have been shown to be a useful tool for the computational prediction of material properties,[3] condensed matter[4] and exploring the dynamical behavior and interactions in huge biomolecular systems.[5–9] However, there are still particular processes for which the explicit inclusion of fully atomistic details remains a must. In these cases, at least two general strategies may be considered to achieve atomic resolution and time-size scalability: (a) recovering AA spatial information from a CG simulation (back-mapping) or (b) performing multiscale or hybrid AA/CG simulations. The first case involves the reconstruction of the atomic coordinates using the information condensed in the CG beads. Despite many successful examples,[10–18] such an approximation is non-trivial and intrinsically dependent on the CG mapping scheme. This problem arises not only due to the differences in granularity but also because the conformational landscape explored by simplified models does not match, in general, that of the AA representation.

On the other hand, in multiscale methodologies, AA and CG representations may be combined in different ways to provide atomistic details on relatively reduced regions of interest incorporating the effect of the macro/supra molecular environment. They can be divided into two categories:[19] serial schemes, where the information obtained at a certain resolution is used to construct a potential function for another level of description; and parallel schemes, where two or more resolution levels are simultaneously present in the system and interact with each other (see ref. 20 and 21 for recent reviews). A critical point in this case regards the description of the interactions at the AA/CG frontier. The expressions for the potential energy describing hybrid representations are frequently calculated as: $E = E_{AA} + E_{CG} + E_{AA/CG}$, where the terms for the atomistic (E_{AA}), simplified (E_{CG}) and hybrid ($E_{AA/CG}$) parts of the system may be evaluated in different ways.

Institut Pasteur de Montevideo, Mataojo 2020, Montevideo, Uruguay. E-mail: spantano@pasteur.edu.uy; Fax: +598 2522 4185; Tel: +598 2522 0910

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c1cp21248f

The form of E_{CG} depends upon the used CG representation and the $E_{AA/CG}$ term has to be tailored to make compatible both levels of description (AA and CG). As the CG schemes are rather arbitrary, the transferability is always an important point. This issue has been addressed in a number of ways combining different simulation approaches for the AA and CG regions of the system.[21–24]

Here we introduce a hybrid (AA/CG) representation for simulating nucleic acids, which can be considered as an extension of a recently published CG scheme for DNA.[25] Usually, in classical molecular dynamics (MD) simulations, the molecular systems are described by a two-body additive Hamiltonian running over all possible atom pairs i, j:

$$U = \sum_{\text{bond}} k_b(r_{ij} - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta(\theta_{ij} - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2}[1 + \cos(n_k\varphi - \gamma_k^{\text{eq}})] + \sum_{\text{non-bond}} \left\{ 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\in r_{ij}} \right\}$$

In the present communication, regions of interest in a given molecule are treated at atomic detail while the rest of the system is simulated at a simplified level. However, all the interactions (corresponding to the terms E_{AA}, E_{CG} and $E_{AA/CG}$) are simultaneously calculated within the same Hamiltonian function (U). This is possible since the interactions in the E_{CG} of our CG approach are described by the same functional form written above for U.[25] Since in our case the sums in the above equation run over all the atoms/beads in the system, the calculation of the interactions (including electrostatics and van der Waals) makes no distinction between the components of the AA, AA/CG or CG regions. In this way no extra terms, particular mixing rules or *ad hoc* shared regions to tie different fragments are required. Only the bonding interactions linking normal atoms with CG beads need to be developed, leaving the existing bonding and non-bonding AA and CG parameters unchanged. We show along the manuscript that this is sufficient to obtain only minor structural and dynamical perturbations in the AA region.

The hybrid AA/CG model for DNA presented here is designed on the basis of a well-established AA force field[26,27] and a CG model developed in our group.[25]

We apply our hybrid model to perform MD simulations embedded in an implicit solvent using the generalized Born model approximation.[28–30] Different double stranded DNA geometries are studied including a single AA/CG border, an AA bubble flanked by two CG segments and a couple of DNA hairpins with sheared GA pairs (Fig. 1). Hybrid MD simulations are compared with analogous calculations at different levels (AA and CG) and against experimental data. We obtained a good performance of the hybrid AA/CG model in all the studied systems with a rather smooth transition for structural and dynamic observables when passing from AA to CG regions. Only slight distortions are noticed at the AA/CG interface and these vanish beyond the first or second base pair step from the interface. After this point, the influence of the CG region on the AA part of the systems, even in terms of the slowly decreasing electrostatic interaction, results negligible. This opens the possibility of including an additional layer at a higher level of complexity (*i.e.* QM/MM) within the AA region.

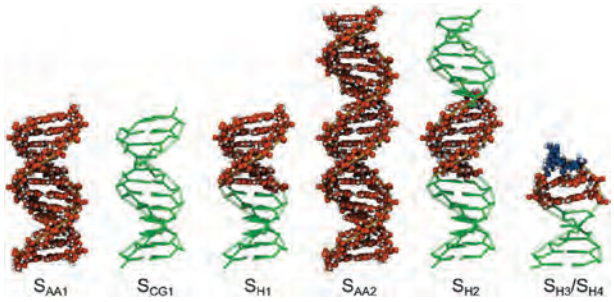

Fig. 1 Structural representation of the studied molecular systems. The CG residues are represented in green sticks while spheres represent AA residues (orange: heavy atoms; white: hydrogen atoms). The transparent tubes connecting the phosphate atoms are only indicative of the backbone. No connectivity is present between consecutive phosphates. Blue residues in the loop region of S_{H3} are changed from thymine to adenine to produce S_{H4}.

Methodology

Molecular systems

Several molecular systems are studied (Fig. 1):

(i) A reference system (S_{AA1}) treated at fully atomistic detail. It consists of the double-stranded Drew–Dickerson dodecamer of DNA,[31] the sequence of which is 5′-d($C_1G_2C_3G_4A_5A_6T_7T_8$-$C_9G_{10}C_{11}G_{12}$)-3′.

(ii) For comparison purposes, the corresponding version of S_{AA1} is simulated at the CG level (S_{CG1}).

(iii) A hybrid version of S_{AA1} (S_{H1}), in which the first half of the structure is treated at the CG level and the second half at the AA level.

(iv) An AA 20-mer double-strand DNA (S_{AA2}) of sequence 5′-d($C_1A_2T_3G_4C_5A_6T_7G_8C_9A_{10}T_{11}G_{12}C_{13}A_{14}T_{15}G_{16}C_{17}A_{18}$-$T_{19}G_{20}$)-3′.

(v) A hybrid version of S_{AA2} (S_{H2}), which is divided into three regions. The CG regions span from base pairs C_1 to T_7 and from A_{14} to G_{20}. The AA segment is placed from base pairs G_8 to C_{13}.

(vi) A single-stranded hairpin DNA (S_{H3}) 5′-d($A_1T_2C_3C_4$-$T_5A_6G_7T_8T_9A_{10}T_{11}A_{12}G_{13}G_{14}A_{15}T_{16}$)-3′ corresponding to a NMR derived structure (PDB code: 1AC7[32]). In this case the CG region extends from residue A_1 to T_5 and from A_{12} to T_{16}, while the region spanning the hairpin loop (bases A_6 to T_{11}) is considered at AA detail.

(vii) A single-stranded DNA hairpin similar to S_{H3} where the two looping thymines (T_8T_9) are changed to adenine to give the sequence 5′-d($A_1T_2C_3C_4T_5A_6G_7A_8A_9A_{10}T_{11}A_{12}G_{13}$-$G_{14}A_{15}T_{16}$)-3′ ($S_{H4}$). The division between CG and AA regions is identical to that of S_{H3}.

Model building

In all cases the model building procedure starts from the Cartesian coordinates of structures containing all the atoms. Systems S_{AA1}, S_{CG1}, S_{H1}, S_{AA2} and S_{H2} are built in the

canonical B-form of DNA[33] using the NAB utility of AMBER10.[34] The atomic coordinates for system S_{H3} are taken from the PDB structure 1AC7,[32] while system S_{H4} is built from S_{H3} by replacing T_8T_9 for A_8A_{10} with the Leap tool of AMBER10.[34] Residues belonging to the CG region are mapped according to the CG scheme published by our group,[25] by simply removing and renaming the corresponding atoms. Residues at the AA region remain unchanged.

Interaction parameters

Molecular dynamics simulations are performed using the AMBER10 package.[34] The parm99 force-field[26] with the correction proposed by Orozco and coworkers for nucleic acids (parmbsc0)[27] is used to represent the AA region while the CG region is described using the parameters reported by Dans *et al.*[25] including its latest modification.[35] The parameterization of the AA/CG interface is done to optimize the fitting to the canonical B-form of DNA. No atoms are removed/changed neither in the AA nor the CG standard residues. The bonding parameters developed in this work are reported in Fig. 2 and Table 1. The hydrogen atoms at the AA region of systems S_{H3} and S_{H4} are replaced by deuterium to match the conditions of the NMR experimental protocol.[32]

Molecular dynamics simulations

The initial structures of each system undergo 1500 steps of energy minimization prior to the simulation. The MD protocol consists of two heating steps of 0.25 ns each, during which the temperature is risen from 0 K to 100 K and then from 100 K to 298 K. After that, a 10 ns equilibration phase is carried out. A reference temperature of 298 K is set by coupling the system to a Langevin thermostat[36] with a friction constant of 50 ps^{-1}. The random seed generator of the stochastic force is changed every restart of the simulation (each 100 ns) to avoid quasi-periodic oscillations.[37] All bonds involving hydrogen atoms within the AA region are restrained using the SHAKE algorithm.[38] The integration step is set to 2 fs for systems S_{AA1}, S_{H1}, S_{AA2} and S_{H2} while the presence of deuterium in systems S_{H3} and S_{H4} allows for a time step of 4 fs. The time step for system S_{CG1} is set to 20 fs. Production runs of 100 ns

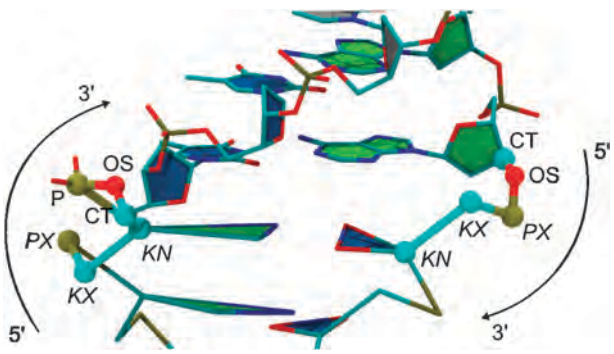

Fig. 2 Molecular representation of the AA/CG interface. AA and CG nucleobases are represented in thin sticks. Atoms and links involved in the interface parameterization are presented as spheres connected with thick sticks (see also Table 1). Atom types in italic characters correspond to the CG force field.[25] The strand directionality is also included. Hydrogen atoms are omitted for clarity.

Table 1 Bonding parameters at the AA/CG frontier

Bond parameters		
	k_b/kcal mol^{-1} $Å^{-2}$	r_{eq}/Å
KN[a]–P	25.0	3.67
OS–PX	100.0	1.59
Angle parameters		
	k_θ/kcal mol^{-1} rad^{-2}	θ_{eq}/°
KN–P–OS	75.0	63.58
KX–KN–P	75.0	96.0
OS–PX–KX	75.0	81.32
CT–OS–PX	75.0	120.0
Dihedral parameters		
	V_1/kcal mol^{-1}	n_1	γ_i^{eq}/°
PX–KX–KN–P	10.0	8	−153.2
KN–P–OS–O2	10.0	8	−140.0
CT–OS–PX–KX	10.0	4	75.6
KN–P–OS–CT	10.0	8	−137.6

[a] KN = KA, KT, KC or KG.

are performed for systems S_{AA1}, S_{CG1}, S_{H1}, S_{AA2} and S_{H2}, while 3 μs are performed for systems S_{H3} and S_{H4}. Snapshots are recorded every 20 ps for analysis. To avoid the possible fraying of the helix ends, loose harmonic restraints (3.0 kcal mol^{-1} $Å^{-2}$) are added to preserve the Watson–Crick hydrogen bonds of the capping base pairs.

Hydration and ionic strength effects are implicitly taken into account using the generalized Born model for implicit solvation.[28–30] The maximum distance between atom pairs considered in the pair wise summation involved in calculating the effective Born radii is set to 10 Å. Non-bonded interactions between 8 Å and 18 Å (slowly-varying terms) are evaluated every 2 integration steps. The salt concentration is set to 0.15 M.

Analysis of the trajectories

Recorded trajectories of hybrid systems (S_{H1} and S_{H2}) are back-mapped to AA coordinates following the procedure defined by us.[25] In this way the comparison is always made among systems having all the atoms, unless otherwise stated. Helical parameters (rise, twist, roll, slide, shift and stretch) are calculated with the software Curves+.[39] In order to eliminate rotational and translational movements all trajectories are fitted to their corresponding canonical B DNA form. Covariance matrix calculation and principal component analysis are performed using the GROMACS utilities G_COVAR and G_ANAEIG.[40] The likeness between the different simulation schemes is estimated by computing the following similarity index (SI):

$$SI = 1 - \sqrt{\left[\mathrm{tr}\left(\sqrt{\frac{M1}{\mathrm{tr}(M1)}} - \sqrt{\frac{M2}{\mathrm{tr}(M2)}}\right)\right]^2}$$

where $M1$ and $M2$ are the two covariance matrices and tr( ) is the trace of the matrix. The SI ranges from one for identical matrices to zero when the sampled subspaces are orthogonal.

Electrostatic potential is calculated on canonical conformers of S_{AA2} and S_{H2} with APBS[41] using a cubic grid of 120 Å per

side with 10 points per Å^2. The potential is forced to converge to zero at the boundaries. The electrostatic potential of both systems is subtracted and the isosurfaces of the difference are drawn at ±10 mV.

Structural comparison of S_{H3} and S_{H4} against each of the 10 NMR structures (PDB code: 1AC7) is done by fitting each trajectory to the corresponding reference and calculating the RMSD on phosphate atoms. The stability of the G_7A_{10} base pair is evaluated by the percentage of hydrogen bond occupancy during the trajectories. The criteria for defining the existence of hydrogen bonds are a donor–acceptor distance below 3.5 Å and donor–acceptor–hydrogen angle smaller than 30°. The loop's structural variability is assessed by clustering the structures of the trajectories into different conformational motifs:[42] type I requires the stacking interaction between the first two or three nucleotides in the loop (G_7, T_8/A_8 and T_9/A_9 in our case). In type II, the base at the second position (T_8/A_8) folds into the minor groove while the third residue in the loop (T_9/A_9) stacks onto the two nucleotides of the duplex stem (*i.e.*, G_7 and A_{10}). In type III there is a continuous stacking of the last three loop bases (T_8/A_8 T_9/A_9 and A_{10}). This conformational arrangement is only found in RNA. Therefore, it is not considered here. To account for the remaining configurations observed during the simulations we define two additional categories: (i) unclassified states correspond to conformations in which there are stacking interactions but the pattern does not correspond to either type I or II; and (ii) disordered states, in which there are no stacking interactions between loop bases. The stacking pattern among bases from G_7 to A_{10} is used to define each motif. Two bases are said to form stacking interaction if the angle between nucleobase planes was greater than 150° or lower than 30° and the distance from their geometric centers was below 5.5 Å. Atoms used to define a nucleobase plane are C8, N1 and C2 for purines and N1, N3 and C5 for pyrimidines. The nucleobase geometric center is calculated using atoms N1, C6, C5, N7, C8, N9, C4, N3 and C2 in purines and N3, C4, C5, C6, N1 and C2 in pyrimidines. Both measurements are performed with the Carnal module of AMBER.[34] The temporal occurrence of type I and type II motifs is also analyzed and representative snapshots from both configurations are taken.

Molecular drawings were performed with VMD 1.8.7.[43]

Results and discussion

MD simulations are performed for a number of double helical DNA molecules at different levels of detail. However, taking profit of the back-mapping capabilities of our CG representation,[25] the analysis is performed on back-mapped trajectories containing all the atoms, unless otherwise stated.

In each of the cases the simulation time is chosen to ensure a proper relaxation of the systems. This is assessed by a low value of the cosine content, which ranges from 0.1 to 0.4, indicating a good convergence within the simulated time.

Comparison between AA, CG and AA/CG representations

As a first test case, we focus on the sequence 5′-d($C_1G_2C_3$-$G_4A_5A_6T_7T_8C_9G_{10}C_{11}G_{12}$)-3′, which corresponds to the Drew–Dickerson dodecamer.[31] Three systems at different levels of description are built in the canonical B form: S_{AA1} (AA), S_{CG1} (CG) and S_{H1} (hybrid AA/CG, Fig. 1). Results are divided into structural and dynamical properties to organize the discussion.

The structural quality of a DNA polymer can be assessed by inspection of its helical parameters. Fig. 3 shows the measurement of rise, twist, roll, slide, shift and stretch from the simulation of S_{AA1} and back-mapped S_{H1}. Averaged crystallographic data taken from ref. 44 and canonical values measured from A and B forms of DNA[33] are also included as a reference. Since the results of atomistic and CG simulations have already been published,[25,35,44,45] we concentrate on the most salient features regarding the hybrid interface.

The rise values of S_{H1} match both experimental data and those from the S_{AA1} simulation (Fig. 3A). The main deviation can be observed at the step 6, *i.e.*, the AA/CG border. However, the average values differ by less than 0.5 Å from those of AA and experimental data with a significant overlap in the standard deviations. This difference is well within the range of variation of both reference data sets, which is below 1 Å. Notably, this relatively small deviation at the frontier does not propagate to the neighboring steps, suggesting that the quality of the simulation in both regions of the molecular system is not worsened by the introduction of the AA/CG border.

The twist measured for the DNA dodecamer is also in good agreement with the reference values (Fig. 3B). The most salient feature within this set of points corresponds to the marked difference with respect to the experimental data obtained for the T_8C_9 step at position 8. In this case, a significant deviation from the crystallographic values is present in the AA simulation. This deviation is also found in S_{H1}, which globally follows the same tendency of the atomistic MD.

The roll values follow different trends at each region of system S_{H1} (Fig. 3C). The CG region (steps 2 to 5) displays negative roll values (around −2°) with small standard deviations. This can be expected as the CG model was developed to reproduce the structural determinants of a B form of DNA. In contrast, the AA region (steps 7 to 10 of S_{H1}) and the atomistic simulation (S_{AA1}) slightly deviate from the B form visiting more positive values (Fig. 3C). However, both families of conformations are compatible with the experimental data within the standard deviations. The opposite tendency in the roll at both regions generates a small perturbation resulting in a few degrees of separation from the canonical value and an increase in the standard deviation at the interface. In contrast to the results from rise and twist, the effect of the AA/CG frontier seems to propagate up to the first neighboring step in the atomistic region of S_{H1}, which displays also negative values. After that point the agreement between S_{AA1} and S_{H1} is recovered.

The results obtained for the slide parameter measured on S_{H1} follow very well those of S_{AA1} (Fig. 3D). Although deviations from experimental data are present at steps 2, 3, 9 and 10, both simulation schemes provide similar results even at the AA/CG interface.

The shift parameter is also well comparable with the experimental and atomistic ones. Again, the major distortion is limited to the bordering step with no sensible influence on any of the neighboring base pairs (Fig. 3E).

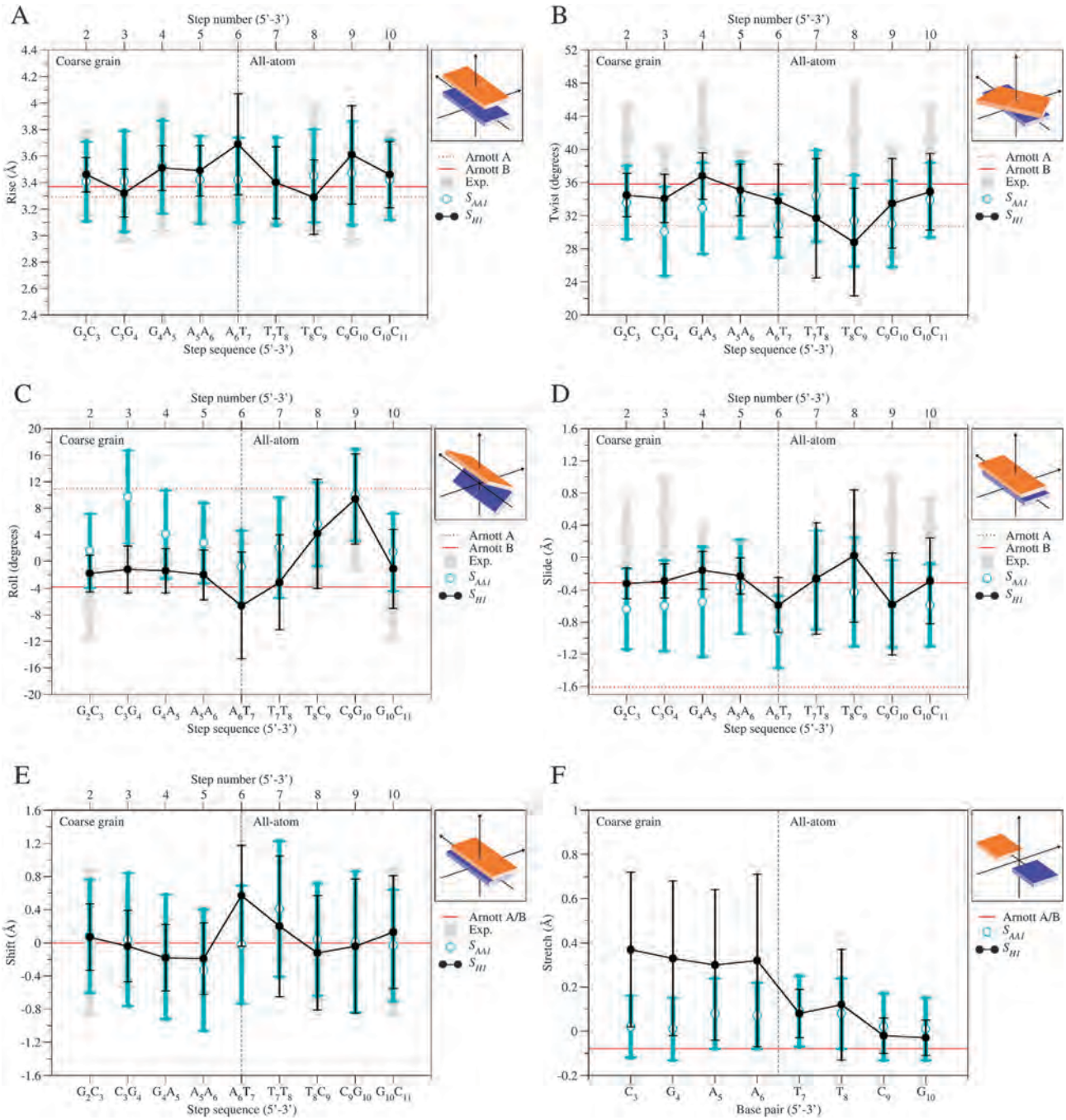

Fig. 3 Comparison between experimental, canonical and simulated helical parameters. (A) rise, (B) twist, (C) roll, (D) slide, (E) shift and (F) stretch. The vertical dashed line indicates the AA/CG frontier in S_{H1}. The canonical values for A and B DNA forms according to Arnott *et al.*[33] are indicated with dashed and continuous red lines, respectively. Filled and open circles represent data for systems S_{H1} and S_{AA1}, respectively. Squares correspond to averaged experimental data taken from ref. 44. Standard deviations are reported as error bars. Analogous data for system S_{CG1} have been previously reported in ref. 25 and 45 and omitted here for the sake of brevity.

The last analyzed helical parameter, the stretch, does not involve consecutive base pair interactions but base–base pairing within the same plane. In this case, the CG region of S_{H1} shows a nearly constant off set of about 0.5 Å with respect to the reference data. This discrepancy seems to be intrinsic to the CG model and can be ascribed to the absence of hydrogen atoms, which contributes to reduce the strength and directionality of the electrostatic interactions. However, only marginal effects are observed in the AA region since already the first AA base pair in S_{H1} nicely matches the stretch value of S_{AA1} (Fig. 3F).

Taken as a whole, the averaged helical parameters suggest a good reproduction of the geometrical features. Therefore, we turn our attention to the analysis of the dynamic behavior of the different representations and the likeness of the conformational spaces sampled by the studied models. To assess the influence

of the CG region on the dynamics of the AA part (and *vice versa*) we calculate similarity indexes (SI) between covariance matrices along the trajectories of AA (S_{AA1}), CG (S_{CG1}) and hybrid AA/CG (S_{H1}) systems. This is done considering the entire molecule as well as separating different segments along the molecule in different trajectories. Similarity indexes are calculated in three different ways: (i) using all the atoms including hydrogen from the atomistic and back-mapped trajectories; (ii) using only the phosphate atoms (*i.e.*, using only one atom per residue) from the atomistic and back-mapped trajectories; and (iii) using only the positions of the atoms corresponding to the CG model (*i.e.*, using six atoms per residue) from the atomistic and *not* back-mapped trajectories. In each case we use the entire set of eigenvectors for the calculation of the corresponding SI. A unitary SI is expected for identical covariance matrices, while a null value is expected for orthogonal matrices.

To establish a reference level for the comparison we take profit of the palindromic character of our test case DNA. Along the S_{AA1} simulation we first separate the trajectories of both palindromic halves (base pairs 1 to 6 and 7 to 12), then we calculate the covariance matrix for each of the halves. From there, we compute the SI between both covariance matrices (see Methodology). We obtained a similarity of 0.87 for all the atoms and 0.91 if we consider only the phosphate atoms (Table 2). This indicates that despite the system's symmetry, some intrinsic variability exists already in the AA model. This may be taken as the maximum similarity value we can expect to find when comparing trajectories at CG or AA levels.

Calculation of the covariance matrices on the entire molecules for S_{AA1} and S_{CG1} using all the back-mapped atoms gives a value of 0.64, reaching 0.72 considering only the phosphate moieties. In general, an increment in the similarity is observed when considering only the phosphate atoms in all the comparisons made between CG and AA segments. This suggests that the global conformational space sampled by both simulation schemes is fairly similar at the backbone level. However, the faster dynamics of the base moieties show more pronounced differences, most likely due to the intrinsic loss of degrees of freedom in the CG scheme that cannot be retrieved even upon back-mapping.

The SI between the entire S_{AA1} and S_{H1} scores 0.66, which points to a sizeable overlap between both sampled subspaces. Comparable values (0.63) are obtained considering only the segments comprising base pairs 1 to 4, 1 to 5 or 1 to 6 (Table 2). The nearly constant results obtained considering different segment lengths indicate that the dynamics of the CG part of the hybrid system is not modified by the closeness to the AA interface. The SI of atomistic regions of S_{H1} with their corresponding segment in S_{AA1} gives a lower value than that obtained in the calculation using both palindromic halves of S_{AA1} (0.73 *vs.* 0.87, Table 2), pointing to a minor change in the conformational space sampled by both regions. In this case, however, a slight increment in the SI is observed in going from base pairs 7 to 12, 8 to 12 and 9 to 12. This could be indicative of some small influence of the CG border on the AA dynamics. The comparison between S_{H1} and S_{CG1} shows an opposite trend. A high SI is found considering the base pairs 1 to 6

Table 2 SI calculated between the covariance matrices of systems S_{AA1}, S_{CG1} and S_{H1}

		S_{H1}							S_{AA1}	S_{CG1}
	Base pair	1–4	1–5	1–6[a]	7–12[b]	8–12	9–12	1–12[c]	1–6	1–12
S_{AA1}	1–4	0.63								
		0.70								
		0.66								
	1–5		0.63							
			0.69							
			0.64							
	1–6			0.63						
				0.69						
				0.64						
	7–12				0.73				0.87	
					0.78				*0.91*	
					0.77				**0.89**	
	8–12					0.74				
						0.80				
						0.78				
	9–12						0.76			
							0.82			
							0.78			
	1–12							0.66		0.64
								0.78		*0.72*
								0.66		**0.64**
S_{CG1}	1–6			0.77						
				0.76						
				0.77						
	7–12				0.60					
					0.69					
					0.61					
	1–12							0.63		
								0.70		
								0.65		

[a] CG region. [b] AA region. [c] Whole molecule. Values in italic correspond to SI calculation considering only the phosphate atoms, while bold characters correspond to the same calculation performed using six atoms per nucleobase without back-mapping the trajectories.

within the CG region of S_{H1} (0.77) while a lower similarity is calculated at the AA region (0.60, Table 2). Both observations support the idea that the CG region of S_{H1} behaves alike the pure CG model (S_{CG1}), while the AA region of S_{H1} explores a landscape more similar to that of the pure AA model (S_{AA1}).

To complete the characterization of the dynamical behavior of the hybrid model, we study the principal components governing the movement of S_{AA1} and S_{H1}. As the most relevant conformational changes in DNA are associated to the backbone the analysis is performed on the phosphate atoms. This ensures that we capture the main distortions of the polymer during the simulation.

On the other hand, the calculations using only the positions of the atoms corresponding to the CG model without back-mapping show essentially the same results (Table 2). This suggests that, if some atomistic dependence is introduced on the CG region by the back-mapping procedure, it is rather negligible.

The eigenvalues corresponding to the essential eigenvectors from the diagonalized covariance matrices present a similar profile on both systems. An abrupt decrease is seen after the third eigenvector. In fact the cumulative sum of the first three is enough to explain more than 60% of the space sampled by both S_{AA1} and S_{H1} (Fig. 4A). Therefore, for the sake of simplicity the analysis is restricted to these three eigenvectors.

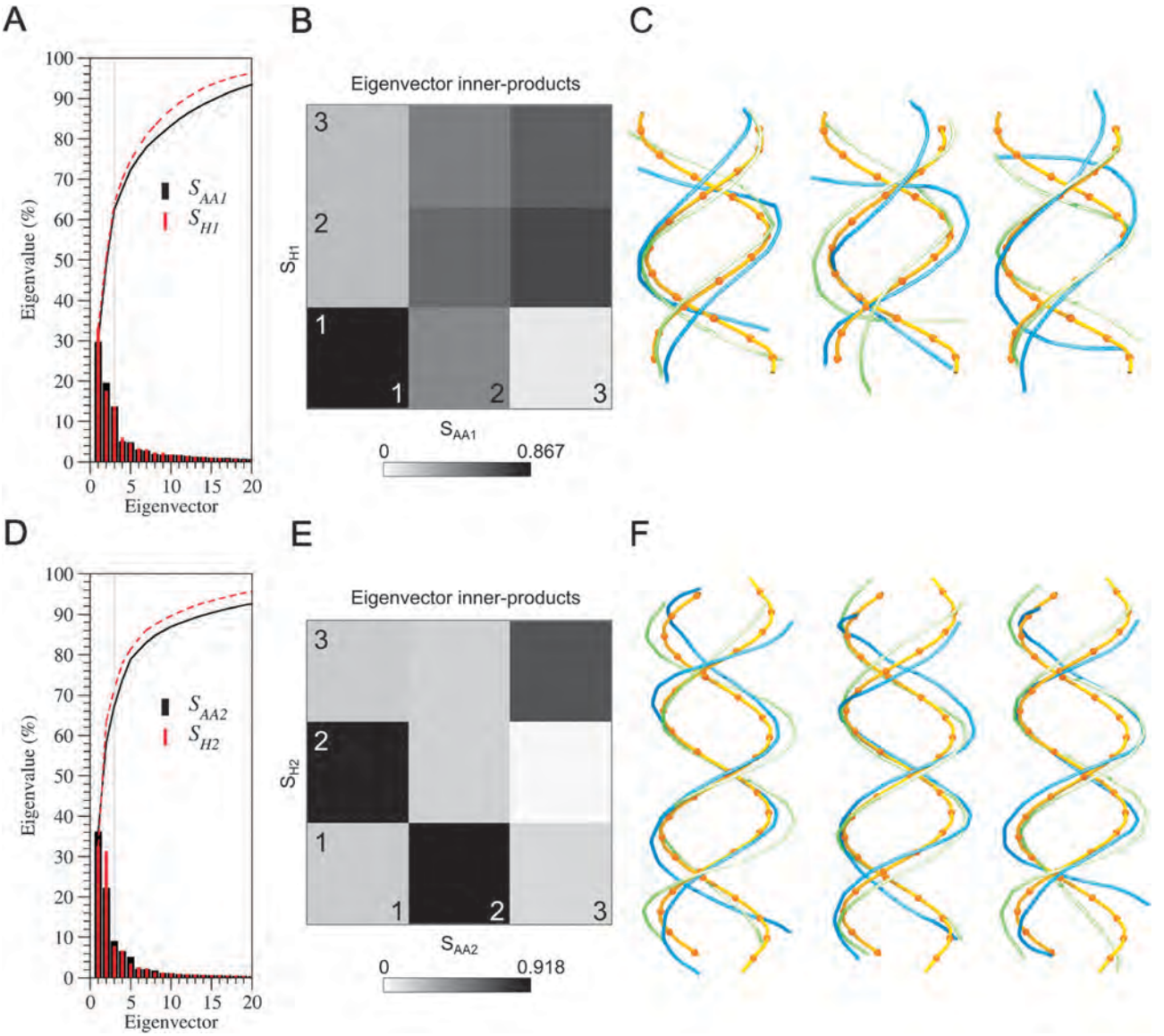

Fig. 4 Principal component analysis on phosphate atoms. (A) The first 20 eigenvectors (out of 66) are plotted against their eigenvalue number. The heights of the bars represent their percentage relative to the trace of the eigenvalue matrix. Black and red colors indicate S_{AA1} and S_{H1}, respectively. Lines (solid or dashed) correspond to the cumulative sum over the preceding eigenvectors for S_{AA1} and S_{H1}, respectively. (B) Matrix of inner-products between the first three eigenvectors of systems S_{AA1} and S_{H1}. (C) Blue and green tubes correspond to the structural representations of extreme projection of the first (left), second (middle) and third (right) eigenvectors of S_{AA1} on the real space. The canonical B form is included as a reference, in which yellow tubes and orange spheres represent the DNA backbone and the phosphate atoms, respectively. (D to F) Same as A to C for systems S_{AA2} (black bars, solid lines) and S_{H2} (red bars, dashed lines). Notice that for systems S_{AA2} and S_{H2} the total number of eigenvectors is 114.

To rationalize the overlap of these vectors on the 3D space trajectories of both systems we calculated the pair-wise inner-product between them. The higher the value of the inner-product, the higher the superposition in the components of the motion described by that particular pair of vectors. Calculation of the inner-product between these six eigenvectors (three for each system) evidences a good degree of superposition in the diagonal elements. The agreement is particularly higher between the two first eigenvalues (Fig. 4B), which corresponds to a bending mode (Fig. 4C, left). However, a certain degree of mixing between the second and third modes can be deduced from the higher inner-product values. Analysis of the extreme projections on the atomistic trajectory suggests that the second and third modes can be mainly ascribed to twisting and twisting/bending movements for the second and third eigenvectors, respectively (Fig. 4C, middle and right).

Analysis of an AA "bubble" within a CG context

The previously described molecular systems constitute a suitable test platform for our hybrid scheme. However, a more interesting and practical example of application would comprise a limited portion of AA residues surrounded by a considerably longer CG segments. Hence, we address the

simulation of a 20-mer DNA sequence (S_{H2}), which consists of an AA bubble of six base pairs placed between two CG regions of seven base pairs each (Fig. 1). As a benchmark the full atomistic model of the system (S_{AA2}) is also studied.

Briefly, the analysis of helical parameters for S_{AA2} and S_{H2} reveals the same trends observed for S_{AA1} and S_{H1}, *i.e.*, the rather small perturbations observed at the AA/CG border vanish after one base pair from the interface (Fig. S1, ESI†). Likewise, a good reproduction is retrieved for the dynamic behavior. Indeed, the SI between the trajectories of S_{AA2} and S_{H2} suggests a high degree of superposition with a value of 0.69 (0.70 for phosphates) considering the entire molecular systems. This value increases reaching a SI of 0.82 (0.83 for phosphates) when calculated only on the six central pair bases simulated at the atomistic level. These results underline the good performance of the hybrid model within the AA region as well as the global behavior of the system.

The longer size of systems S_{AA2} and S_{H2} offers a good opportunity to analyze in more detail the principal components, which are dominant in the dynamics. Fig. 4D shows that the first three eigenvectors associated to S_{AA2} and S_{H2} explain $\sim$70% of the motion. In S_{AA2} the first eigenvector stands out from the others (representing $\sim$36% of the variance), while first and second eigenvectors of S_{H2} have an equivalent weight ($\sim$32%). This observation acquires relevance when analyzing the inner-products matrix. Direct comparison between the first two eigenvectors of S_{AA1} and S_{H2} suggests that the spaces explored by them are different (Fig. 4E). However, when performing a cross comparison (*i.e.*, eigenvector 1 of S_{AA2} against eigenvector 2 of S_{H2} and *vice versa*), we find significantly higher values in the off-diagonal elements. This indicates that the movement associated to eigenvector 1 in S_{AA2} is represented by eigenvector 2 in S_{H2} and *vice versa*. Visualization of the extreme projections of these two first vectors in S_{AA2} shows that they correspond to nearly perpendicular bending modes in the double helical DNA fragment (Fig. 4F, left and middle). We can conclude that, despite the discrepancy in the eigenvector modules, the same movements are represented along both trajectories with slightly different weights.

In contrast with the behavior of the first two vectors, we retrieve a good overlap between both third eigenvectors in the trajectories of S_{AA2} and S_{H2}, as evidenced by the high value of the corresponding inner-product (Fig. 4E). This vector can be identified as a twisting movement (Fig. 4F, right), which is nearly equally represented in both systems (Fig. 4D).

As already stated, the scheme presented here is based on the introduction of bonding parameters linking the AA/CG interface. Besides the minor perturbations in the structural and dynamical features, it is also important to evaluate the possible spurious effects arising from a misbalance in the non-bonding interactions. In particular we concentrate on the electrostatic potential, which has the longest relaxation distance. With this aim we calculated the electrostatic potential generated by the canonical form of S_{AA2} and S_{H2} on a grid surrounding each system. Then we computed the difference between the two electrostatic potential grids. As shown in Fig. 5A, appreciable differences exist only within or near the CG region. The isosurfaces are drawn at $\pm$10 mV, which is a tiny figure if we consider that fluctuations across a box of pure SPC water are, on a temporal average, in the order of 4 mV.[35] If we consider 10 mV as the maximum acceptable perturbation on the electrostatic potential we should conclude that at least two atomistic base-pair steps would be needed to buffer the influence of the CG region on the AA segment. The close up in Fig. 5B and C shows that the regions where the differences are more sensible are those of the sugar rings (minor groove) at the rim of the second base-pair step from the hybrid interface, while the corresponding base moieties feel no difference in electrostatic potential. Although the differences in the electrostatic potential propagate up to a couple of steps, these differences have seemingly no effects as the perturbations in dynamical and structural features introduced by the CG regions on the AA bubble are noticeable, at most, up to the first step (Fig. S1, ESI†).

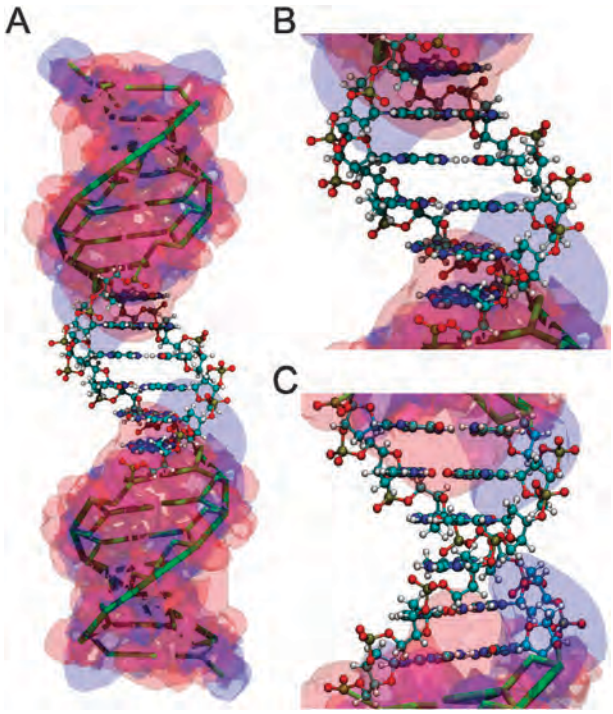

Fig. 5 Differences in the electrostatic potential. The difference between electrostatic potential grids of AA (S_{AA2}) and AA/CG (S_{H2}) schemes is calculated and the results are mapped in the 3D space. The canonical structure of system S_{H2} is included as a reference. Positive (blue) and negative (red) isosurfaces are drawn at values of +10 and −10 mV respectively. (A) Global view. (B) Close up on the AA region looking into the mayor groove. (C) Same as B but rotated 90° around the principal DNA axis.

The data presented in Fig. 3 and Fig. S1 (ESI†) seem to indicate that the introduction of additional interaction (perhaps non-Hamiltonian) terms could further reduce the discrepancies observed with the fully atomistic simulation in terms of helical parameters. In particular, van der Waals interactions at the hybrid interface are poorly reproduced, as a considerably reduced number of effective beads in the CG residue are opposed to the atoms in the AA base. The perturbations introduced by this unbalance extend up to the first or second base pair from the interface. On the other hand, electrostatic perturbations extend up to the second base pair (Fig. 5).

Hence, although incorporation of additional interactions would likely improve the behavior of the AA/CG link, it would not reduce the number of nucleotides needed to smooth the perturbations in the AA segment since they are dominated by long-range electrostatic interactions.

Study of two hairpin systems

As an example of application of our hybrid AA/CG strategy we undertook the study of two DNA hairpins (S_{H3} and S_{H4}, Fig. 1). This molecular architecture includes a double helical DNA stem capped by a tetranucleotide loop, which is stabilized by the formation of a sheared GA pair. This non-Watson–Crick pairing is not well reproduced using our CG model as the interaction points are not present in our simplified scheme,[25] highlighting the usefulness of the hybrid scheme.

DNA hairpins are principally found in prokaryotes and their viruses and play important biological roles in different kinds of organisms during replication, recombination and transcription.[46] As some proteins can directly recognize and bind DNA hairpins in a sequence dependent way, their structure and dynamics at the loop region have a biological relevance.[46] In particular, the hairpin studied here is related to telomeric and centromeric structures and has been solved by NMR spectroscopy.[32] Besides the original structure containing a $G_7T_8T_9A_{10}$ tetraloop (S_{H3}), we also simulated a modified version of the hairpin containing adenines at positions 8 and 9 (S_{H4}), which are not supposed to alter the structural stability of the loop despite the significantly different chemical characters.

Structural comparison between the ensemble of NMR structures and the corresponding MD trajectories (S_{H3} and S_{H4}) gives an overview on the landscapes explored by the models. The RMSD values of the backbone with respect to any of the NMR structures range from 1.0 to 5.0 Å (Fig. 6). Over 70% of the simulation time the models sample conformations close to at least one of the experimental conformations (RMSD below 2.0 Å). While nearly 30% of the remaining time they explore conformations between 2.0 and 3.0 Å of RMSD. Deviations beyond 3.0 Å are extremely rare during the simulation (less than 3% of occurrence), pointing to a good global agreement with the experimental data. During the simulation the hairpin explores different conformations approaching to different NMR reference structures with time periods that vary from tens of nanoseconds to almost one microsecond as evidenced by the alternation of different colors in Fig. 6.

A relevant interaction in both tetraloops (bases G_7 T_8/A_8 T_9/A_9 A_{10}) corresponds to the contact between G_7 and A_{10}. These two nucleotides interact though non-standard Watson–Crick interactions (sheared $G_{(anti)}A_{(anti)}$ pair) while nucleotides at positions 8 and 9 may form different types of stacking interactions.[32] The G_7A_{10} pair is stabilized by the formation of two hydrogen bonds in the *anti–anti* conformation involving the pairs N7 (G_7) – N2 (A_{10}) and N6 (A_{10}) – N3 (G_7) as hydrogen donors and acceptors, respectively. The simultaneous presence of both hydrogen bonds is higher than 80% during both simulations and at least one of them is always present, underlining the relevance of the G_7 and A_{10} pair for the structural stability of the hairpin.

The two central bases within the tetraloop are not involved in stable hydrogen bonds. However, they engage stacking interactions, which may be determinant for the conformation of the loop. While these loops can be classified into three types,[42] the third class (Type III) is not considered here, as it is only present in RNA hairpins. Moreover, we introduced two

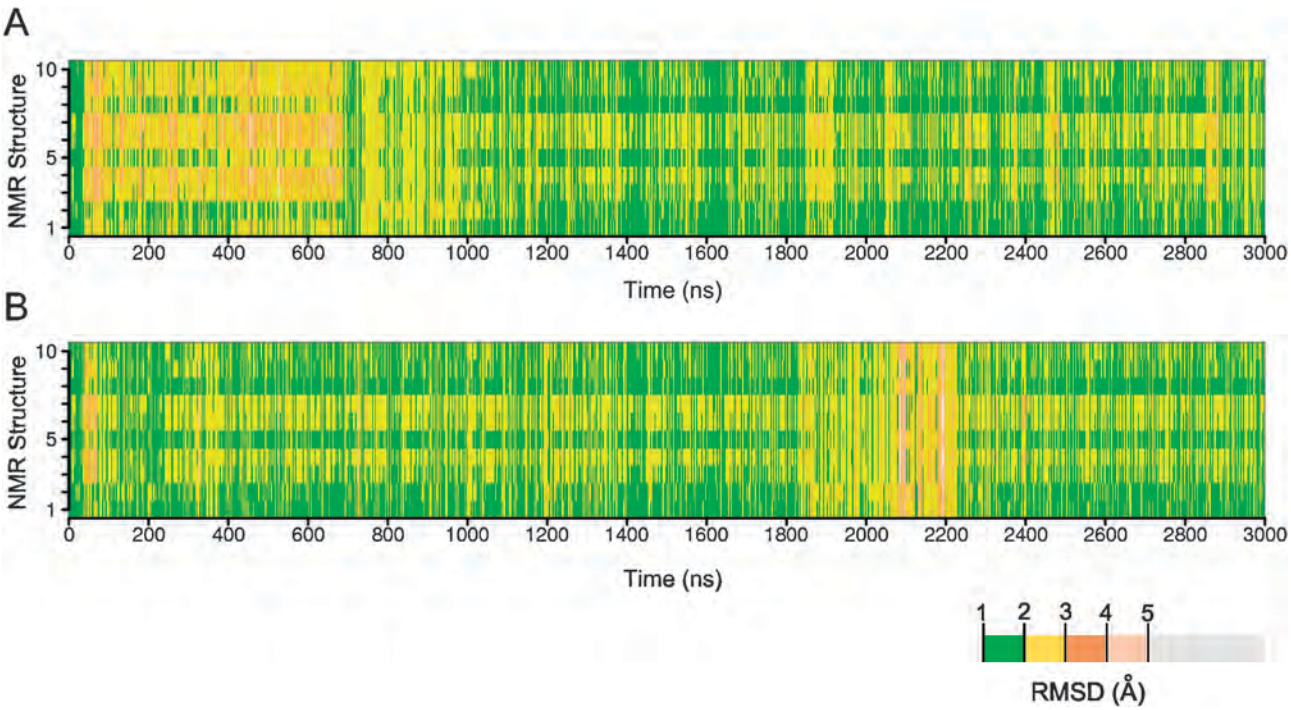

Fig. 6 Conformational behavior of the DNA hairpin. (A) Each colored row corresponds to a RMSD of the phosphate atoms during the MD trajectory of S_{H3} using as reference each of the ten conformers derived by NMR (PDB code: 1AC7). The instantaneous RMSD values are depicted according to the color scale at the bottom of the figure. A sampling time of 20 ps was used. (B) Same as A for S_{H4}.

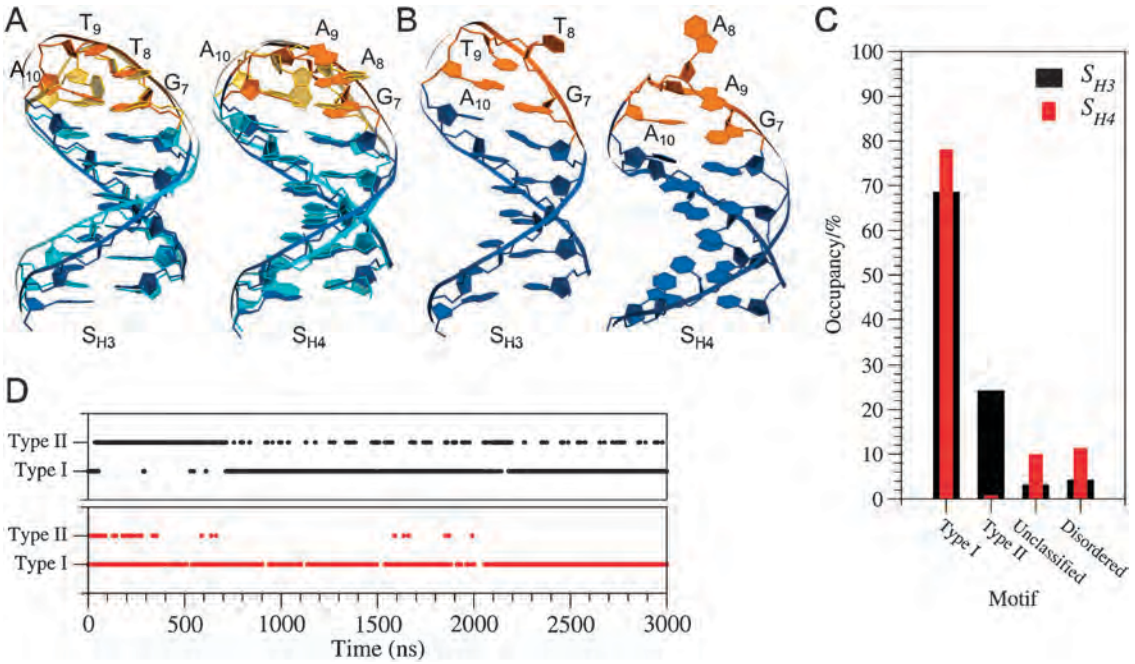

Fig. 7 Hairpin dynamics at the loop region. Representative structures of (A) type I and (B) type II motifs. The lighter models superimposed on A represent alternative conformers fulfilling the definition of a type I loop. (C) Occupancy of loop configurations visited by S_{H3} (black wide bars) and S_{H4} (red thin bars). (D) Temporal occurrence of type I and II motifs along the MD of systems S_{H3} (up, black lines) and S_{H4} (bottom, red lines).

additional categories (unclassified and disordered) to account for other conformations visited during the dynamics but not comprised in the definition of loops type I or II (see Methodology).

According to these definitions, two main populations corresponding to type I are observed in S_{H3} and S_{H4} along the simulations. The first shows a continuous stacking of the nucleotides at positions 7, 8 and 9, while only nucleotides at positions 7 and 8 are involved in the stacking interaction in the second population (Fig. 7A). In addition, we also observe type II conformations where the residue at position 9 stacks indistinctly on either one or both nucleobases involved in the sheared GA pair (Fig. 7B). From the structural point of view, substitution of T_8T_9 in S_{H3} by A_8A_9 in S_{H4} does not introduce significant conformational modifications in any of the loop types. The most populated configurations visited by S_{H3} and S_{H4} correspond to type I (>70% occupancy, Fig. 7C), in agreement with the experimental information.[32] However, the pyrimidine to purine substitution has a clear impact on the loop's dynamics owing to the tendency of purines to form more stable stacking interactions than pyrimidines. In fact, S_{H3} are more prone to visit type II configurations than S_{H4}, whose type I occupancy is 20% higher. Furthermore, the higher steric hindrance of adenines appears to alter the propensity of the loop to visit different conformational states, as the occupancy of unclassified or disordered states is also increased during the S_{H4} simulation (Fig. 7C).

Regardless the relative occupancy of different states, both simulations suggest that the transition from type I to type II is a dynamic process with characteristic transition times near the microsecond (Fig. 7D). For instance, microseconds long conformational dynamics have been recently reported for the apical loop element of the nascent HIV-1 RNA transcripts (TAR[47]). Thus, highlighting the potential utility of hybrid approaches to extend the spatiotemporal scales accessible to computer simulations keeping trace of atomistic information.

Conclusions

We presented here a set of parameters, which straightforwardly allows us to link atomistic and simplified representations of nucleotides in MD simulations. The set of simulations presented here shows that the AA/CG transition is effectively smooth, and in the few cases where perturbations are detected, they converge to atomistic values within the first or second base pair after the interface. This is particularly important for the case of the slowly decaying electrostatic potential. This suggests, as a general rule, that at least two base pairs beyond the region of atomistic interest are needed to soften the (relatively small) perturbations introduced by the CG region in the AA segment.

The agreement resulting from the systematic comparison between atomistic, coarse grain and hybrid representations of a series of systems proposes this strategy as a very promising one.

Our hybrid AA/CG interface for double stranded DNA introduces only a reduced set of linking interactions (two bonds, four angles and four dihedrals, Table 1). This is accomplished without modifying the existing parameterization of the nucleotides at AA and CG levels. Furthermore, the back-mapping capability of our CG model grants the possibility to obtain atomic details for the entire AA/CG system. Using this scheme, the considerably higher performance granted by CG approaches can be straightforwardly complemented to treat non-standard interactions, modifications and complexes at fully atomistic detail. The speed-up resulting from this approach will intrinsically depend on the relative sizes of each of the components of the system (AA and CG). In the present examples, the speed-up

comparing systems S_{AA1} *vs.* S_{H1} (half of the molecule treated at the AA level), resulted of 50%, while comparison of S_{AA2} *vs.* S_{H2} (30% of the system treated at the AA level), give a speed-up of 70%. Moreover, the computer cost to simulating S_{H1} and S_{H2}, which have the same number of AA residues was nearly identical. Underlining the advantage of the hybrid approach to extend either the size and/or time scale accessible to MD simulations. A further acceleration can be expected from the use of multi-timestep approaches setting a longer integration period for the CG region. This possibility, however, has not been explored in this work.

A more ambitious perspective on the potentiality of this scheme can be acquired by considering the successes of QM/MM methods, which opened the possibility to include macromolecular effects (electrostatic and mechanical coupling) in high-level molecular calculations.[48–51] Since the set of parameters presented here for the interface and the CG region are built in together with the classical part of the calculations, it straightforwardly allows to explicitly consider the supramolecular environment at multiscale level (QM/AA/CG) of description.

Acknowledgements

We would like to thank Leonardo Darré for helpful comments on the manuscript. 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). Computer time granted by cluster-FING (http://www.fing.edu.uy/cluster) is also acknowledged. M. R. Machado is beneficiary of a National Fellowship provided by CSIC-UdelaR (Comisión Sectorial de Investigación Científica-Universidad de la República). P. D. Dans and S. Pantano are researchers from the National Scientific Program of ANII (SNI) and from PEDECIBA (Basic Science Development Program of Uruguay).

References

1 M. Cascella and M. Dal Peraro, *Chimia*, 2009, **63**, 14–18.
2 G. A. Voth, in *Coarse-graining of condensed phase and biomolecular systems*, ed. G. A. Voth, CRC Press/Taylor & Francis Group, New York, 2009, pp. 1–455.
3 A. Beaber and W. Gerberich, *Nat. Mater.*, 2010, **9**, 698–699.
4 C. Peter and K. Kremer, *Soft Matter*, 2009, **5**, 4357–4366.
5 G. S. Ayton and G. A. Voth, *Biophys. J.*, 2010, **99**, 2757–2765.
6 M. Durrieu, P. J. Bond, M. S. P. Sansom, R. Lavery and M. Baaden, *ChemPhysChem*, 2009, **10**, 1548–1552.
7 H. J. Risselada and S. J. Marrink, *Phys. Chem. Chem. Phys.*, 2009, **11**, 2056–2067.
8 W. Shinoda, R. DeVane and M. L. Klein, *J. Phys. Chem. B*, 2010, **114**, 6836–6849.
9 Y. Yin, A. Arkhipov and K. Schulten, *Structure (London)*, 2009, **17**, 882–892.
10 A. P. Heath, L. E. Kavraki and C. Clementi, *Proteins: Struct., Funct., Bioinf.*, 2007, **68**, 646–661.
11 L. Chen, H. Qian, Z. Lu, Z. Li and C. Sun, *J. Phys. Chem. B*, 2006, **110**, 24093–24100.
12 B. Hess, S. León, N. van der Vegt and K. Kremer, *Soft Matter*, 2006, **2**, 409–414.
13 G. Santangelo, A. D. Matteo, F. Müller-Plathe and G. Milano, *J. Phys. Chem. B*, 2007, **111**, 2765–2773.
14 P. Liu, Q. Shi, E. Lyman and G. A. Voth, *J. Chem. Phys.*, 2008, **129**, 114103.
15 P. Carbone, H. A. Karimi-Varzaneh and F. Müller-Plathe, *Faraday Discuss.*, 2010, **144**, 25–42; discussion 93–110, 467–481.
16 A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risselada, A. H. De Vries and S. J. Marrink, *J. Comput. Chem.*, 2010, **31**, 1333–1343.
17 P. J. Stansfeld and M. S. P. Sansom, *J. Chem. Theory Comput.*, 2011, **7**, 1157–1166.
18 V. Tozzini, W. Rocchia and J. A. McCammon, *J. Chem. Theory Comput.*, 2006, **2**, 667–673.
19 G. S. Ayton, W. G. Noid and G. A. Voth, *Curr. Opin. Struct. Biol.*, 2007, **17**, 192–198.
20 V. Tozzini, *Acc. Chem. Res.*, 2010, **43**, 220–230.
21 S. O. Nielsen, R. E. Bulo, P. B. Moore and B. Ensing, *Phys. Chem. Chem. Phys.*, 2010, **12**, 12401–12414.
22 M. Neri, C. Anselmi, M. Cascella, A. Maritan and P. Carloni, *Phys. Rev. Lett.*, 2005, **95**, 218102.
23 M. Orsi, M. G. Noro and J. W. Essex, *J. R. Soc., Interface*, 2010, **59**, 826–841.
24 G. S. Ayton, E. Lyman and G. A. Voth, *Faraday Discuss.*, 2010, **144**, 347–357; discussion 445–481.
25 P. D. Dans, A. Zeida, M. R. Machado and S. Pantano, *J. Chem. Theory Comput.*, 2010, **6**, 1711–1725.
26 J. Wang, P. Cieplak and P. A. Kollman, *J. Comput. Chem.*, 2000, **21**, 1049–1074.
27 A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817–3829.
28 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *Chem. Phys. Lett.*, 1995, **246**, 122–129.
29 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem.*, 1996, **100**, 19824–19839.
30 V. Tsui and D. A. Case, *Biopolymers*, 2000, **56**, 275–291.
31 H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. E. Dickerson, *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 2179–2183.
32 M. J. van Dongen, M. M. Mooren, E. F. Willems, G. A. van der Marel, J. H. van Boom, S. S. Wijmenga and C. W. Hilbers, *Nucleic Acids Res.*, 1997, **25**, 1537–1547.
33 S. Arnott, P. J. Campbell-Smith and R. Chandrasekaran, *Nucleic Acids*, in *Handbook of biochemistry and molecular biology*, ed. CRC Press, Cleveland, 3rd edn, 1976, vol. II, pp. 411–422.
34 *AMBER 10*, University of California, San Francisco, 2008.
35 L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera and S. Pantano, *J. Chem. Theory Comput.*, 2010, **6**, 3793–3807.
36 X. Wu and B. R. Brooks, *Chem. Phys. Lett.*, 2003, **381**, 512–518.
37 D. J. Sindhikara, S. Kim, A. F. Voter and A. E. Roitberg, *J. Chem. Theory Comput.*, 2009, **5**, 1624–1631.
38 J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327.
39 R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, *Nucleic Acids Res.*, 2009, **37**, 5917–5929.
40 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
41 N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10037–10041.
42 S. L. Lam and L. M. Chi, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2010, **56**, 289–310.
43 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38, 27, 28.
44 A. Pérez, F. Lankas, F. J. Luque and M. Orozco, *Nucleic Acids Res.*, 2008, **36**, 2379–2394.
45 P. D. Dans, L. Darré, M. R. Machado, A. Zeida and S. Pantano, in *A course on biomolecular simulations*, ed. J. Villà-Freixa, Huygens Editorial, in press.
46 D. Bikard, C. Loot, Z. Baharoglu and D. Mazel, *Microbiol. Mol. Biol. Rev.*, 2010, **74**, 570–588.
47 E. A. Dethoff, A. L. Hansen, C. Musselman, E. D. Watt, I. Andricioaei and H. M. Al-Hashimi, *Biophys. J.*, 2008, **95**, 3906–3915.
48 A. Laio, J. VandeVondele and U. Rothlisberger, *J. Chem. Phys.*, 2002, **116**, 6941.
49 E. Pellegrini and M. J. Field, *J. Phys.Chem. A*, 2002, **106**, 1316–1326.
50 G. D. M. Seabra, R. C. Walker, M. Elstner, D. A. Case and A. E. Roitberg, *J. Phys. Chem. A*, 2007, **111**, 5655–5664.
51 H. L. Woodcock, B. T. Miller, M. Hodoscek, A. Okur, J. D. Larkin, J. W. Ponder and B. R. Brooks, *J. Chem. Theory Comput.*, 2011, **7**, 1208–1219.

Amino Acids (2010) 38:1571–1581
DOI 10.1007/s00726-009-0371-3

ORIGINAL ARTICLE

Isoform-specific determinants in the HP1 binding to histone 3: insights from molecular simulations

Matias R. Machado · Pablo D. Dans · Sergio Pantano

Received: 8 September 2009 / Accepted: 12 October 2009 / Published online: 4 November 2009

Abstract Despite the significant improvements in anti HIV-1 treatment, AIDS remains a lifelong disease due to the impossibility to eradicate the viral reservoir established upon integration of the viral genome. Controlling the epigenetic block imposed by the host cell machinery to the viral transcription may represent a therapeutic alternative to purge the viral reservoir, offering a way to eradicate the infection. Heterochromatin protein 1 (HP1) has been reported to actively participate in the silencing of HIV-1 integrated genome by binding to histone 3 (H3) tail. This interaction is mediated by the Chromodomain of HP1. Nevertheless, the structural features that determine its binding to H3 tail upon post-transductional modifications, such as methylation and phosphorylation as well as isoform-specific effects have not yet been described. We have undertaken the systematic simulation of the Chromodomains of the isoforms beta and gamma of HP1 in complex with the H3 tail methylated at Lys9 in presence/absence of phosphorylation at Ser10. Our results pinpoint isoform-specific electrostatic interactions as important determinants for the stability of the complexes. Characterization of intermolecular contacts between HP1 variants and H3 furnishes new insights on isoform-specific recognition and the effect of phosphorylation.

Keywords Epigenetics · HIV-1 · Transcription · Phosphorylation · Methylation

M. R. Machado · P. D. Dans · S. Pantano (✉)
Institut Pasteur of Montevideo, Mataojo 2020,
11400 Montevideo, Uruguay
e-mail: spantano@pasteur.edu.uy

S. Pantano
IMASL, CONICET, National University of San Luis,
Ejercito de los Andes 950, CP 4700, San Luis, Argentina

Introduction

HIV-1 infection can be effectively controlled by highly active anti-retroviral therapy, with a clear improvement in the life quality of the infected individuals (Clavel and Hance 2004; Martinez-Cajas and Wainberg 2008). However, current therapies cannot cure HIV-1 disease and compliance failures cause rebound of viremia, favouring the evolution of escape variants that are resistant to current drugs (McKinnon et al. 2009). Hence, there is still a need for drugs directed towards different targets, other than those addressed by current therapies, and for treatments that go in the direction of curing the disease by eradicating the infection. Therapeutic targeting of viral post-integration latency is a major goal to attempt HIV-1 eradication (Pierson et al. 2000).

The repression imposed by chromatin is an important factor for maintenance of viral reservoirs and several strategies that selectively activate quiescent proviral genomes with relatively limited effects on the host cell have been proposed (Ylisastigui et al. 2004; Lehrman et al. 2005). In that sense, the understanding of the molecular mechanisms involved in the silencing/repression of the integrated retroviral genome remains largely unknown. It has been hypothesized that heterochromatin machinery and repressive histone marks may play a determining role in chromatin-mediated HIV-1 transcriptional silencing (Marban et al. 2007). It is widely known that histone methylation is involved in heterochromatin assembly and gene silencing (Grewal and Moazed 2003). In particular, Heterochromatin protein 1 (HP1) specifically recognizes histone 3 (H3) methylated at Lys9 (Bannister et al. 2001; Jacobs et al. 2001; Lachner et al. 2001). A consequence of HP1 recruitment is the establishment of a chromatin repressive state that leads to gene silencing (Grewal and

Moazed 2003; Maison and Almouzni 2004). In humans there are three isoforms of HP1 that differ in its nuclear localization (Maison and Almouzni 2004). While HP1 α and β are mainly concentrated at pericentric heterochromatin, HP1γ also localizes to euchromatic sites (Minc et al. 2000; Nielsen et al. 2001). It has been reported that in absence of stimulation, HP1β is present on the HIV-1 promoter together with the non-processive RNAPII and functions as a negative regulator. However, HP1β bound to H3 methylated at Lys9 may be released concurrent with H3 phospho-acetylation, and replaced by HP1γ (Mateescu et al. 2008). This isoform localizes to the HIV-1 promoter but also inside the coding region, together with the processive RNAPII (Nielsen et al. 2001). An independent line of evidence established that HIV-1 reactivation could be achieved after RNA interference against HP1γ in different cellular models, suggesting that targeting only the HP1γ isoform can be sufficient to achieve HIV-1 derepression (du Chéné et al. 2007). Therefore, derepression of chromatin at the HIV-1 integration site by modulating the interaction between H3 trimethylated at Lys9 (H3K9Me3) and HP1 may represent a target for drugs aiming at reactivating the virus from post-integrative latency.

Several reports suggest that the epigenetic mark to release H3K9Me3-mediated repression is the phosphorylation of H3 at Ser10 (H3S10p) (Fischle et al. 2005; Hirota et al. 2005; Johansen and Johansen 2006). However, some controversy remains as it has been proposed that this post-transductional modification can be tolerated by the HP1–H3K9Me3 complex (Mateescu et al. 2004). In addition to H3S10p, acetylation of H3 at Lys14 has also been proposed to abrogate the protein–protein interaction (Mateescu et al. 2004).

HP1 binds to H3K9Me3 through a very conserved folding module called Chromodomain (from CHROmatin MOdifier). This domain is formed by nearly 60 residues folded in a three-strand β-roll ended by a C-terminal α-helix (Fig. 1a, b). This architecture gives place to a binding site where an extended peptide binds forming an anti-parallel β-sheet between strand β1 and residues of the β3-α1 loop, which in absence of the ligand remain unstructured (Ball et al. 1997). The binding pocket contains a critical Triptophan and two Tryrosine/Phenilalanines that coordinate the binding of a three or dimethylated Lysine via a triple cation-π interaction. Affinity is further strengthened by the presence of a highly conserved acidic residue, which confers to the trimethyl lysine unique recognition characteristics (Jacobs and Khorasanizadeh 2002).

The high similarity between HP1 isoforms (Fig. 1a) suggests that subtle differences in the binding domain may determine differential interactions with H3K9Me3. Only a few experimental structures of HP1 Chromodomain in complex with a methylated histone tail are available (Jacobs and Khorasanizadeh 2002; Nielsen et al. 2002), while no structural information is currently known for the Ser10 phosphorylated form of the adduct, nor for the γ isoform.

In this contribution we aim to provide a comparative view into the structural determinants that rule the complexes between H3K9Me3 and isoforms β and γ of HP1, which are involved in HIV-1 promoter transcription/silencing (du Chéné et al. 2007; Mateescu et al. 2008). Moreover, in an effort to provide new insights onto the effect of phosphorylation at Ser10 we constructed structural models of the doubly modified complex (HP1–H3K9Me3S10p). Molecular dynamics (MD) simulations were used to relax the models and evaluate its interactions and stability upon temperature effects.

Our results point to a higher stability of the complex with HP1β compared to that with HP1γ. Ligand detachment upon phosphorylation at Ser10 was not observed (perhaps due to limited sampling time). Modification of this residue seems to be more tolerated in the γ isoform owing to a reduced electrostatic repulsion. This study highlights the influence of the interactions between the N-terminal of the Chromodomain and basic residues at both sides of the trimethylation site. These contacts are isoform specific and could be exploited to increase the selectivity of rationally designed compounds with potential anti HIV-1 activity.

Methods

Molecular systems

We used as template the structure of the mouse HP1β Chromodomain (considered as the receptor) bound to a peptide from H3 dimethylated at lysine 9 (considered as the ligand) (PDB entry 1GUW; Nielsen et al. 2002). It is expected to be identical to the human counterpart (100% of identity within the region used in this work comprising the Chromodomain, i.e. residues 15 to 72 in mouse and human). It is worth to notice that Lys9 in this structure is dimethylated instead of trimethylated. Nevertheless, structural comparison between di- and trimethylated H3 peptides bound to the Drosophila HP1 Chromodomain shows that the only minor differences regarding the coordination of a water molecule by the dimethylammonium moiety (Jacobs and Khorasanizadeh 2002). Therefore, no changes were introduced in the receptor scaffold. Dimethylation present at lysine 4 of the ligand peptide was removed.

Starting from the NMR structure (Nielsen et al. 2002), five systems were built

(i) A complex of the human HP1β bound to the N-terminal 18-mer of H3 trimethylated at Lys9 (H3K9Me3). This

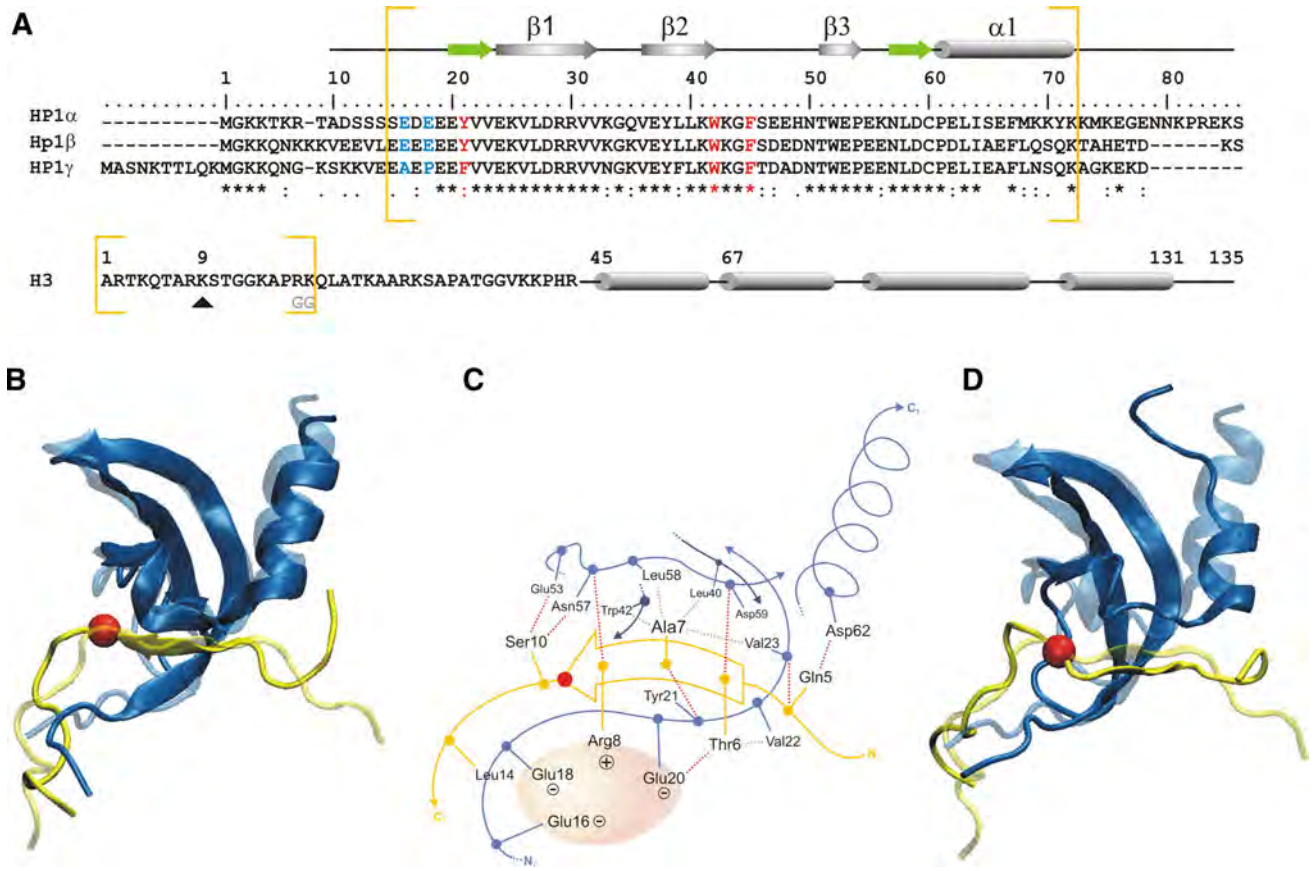

Fig. 1 Molecular systems and interactions. **a** *Top* Sequence alignment of the N-terminal segment of the three human isoforms of HP1. The secondary structure is indicated on the top of the alignment. *Green arrows* represent protein segments which structure approaches to a β-strand upon H3 binding. *Squared brackets* indicate the region comprising the structure of the Chromodomain used in this work. *Red* residues are conserved amino acids constituting the aromatic cage for cation-π interaction. *Blue* residues depict sequence changes in β and γ isoforms at the N-terminal region of the model. Numbering corresponds to the beta isoform. *Bottom* Schematic representation of histone H3. The K9Me3 residue is pointed by a *black triangle*. In this sequence *brackets* define the sequence region included in the model. Note that residues Arg17 and Lys18 were actually replaced by Gly in the NMR template structure (1GUW) and conserved in this study. **b** Least RMSD fit of the initial (translucent) and final conformers of the HP1β–H3 complex. *Blue* and *yellow* parts of the figure represent HP1β and the H3K9Me3 peptide, respectively. The C carbon of K9Me3 is indicated as a *red ball* for reference. **c** Schematic representation of the interactions discussed in the text. *Red* and *grey dashed lines* indicate Hbond and hydrophobic interactions, respectively. The *red oval* represents a region of negative potential generated by the acidic N-terminal of the Chromodomain. **d** Same as **b**, for the complex with the gamma isoform

corresponds to the first model of the NMR family (best representative conformer in the ensemble of 25 structures).

(ii) A complex of the human HP1γ bound to the same peptide of System *i*. This was obtained by mutating the residues according to the sequence alignment shown in Fig. 1a. Point mutations were introduced removing the side chains of the mutating residues and adding the corresponding atoms in its canonical conformation. Possible clashes were relaxed by global energy minimization (see below).

(iii) A complex of the human HP1β bound to the H3K9Me3 in which *Glu16* and *18* were mutated to *Ala* and *Pro*, respectively. These point mutations were introduced on the final minimized conformer of system *i* (after 60 ns, see below) following the same procedure used for system *ii*.

(iv) A complex of the human HP1β Chromodomain bound to the H3K9Me3 phosphorylated at Ser10 (H3K9Me3S10p).

(v) Same as *iv* but for the HP1γ isoform. These models were then used as initial coordinates for MD simulations.

Molecular dynamic simulations

All simulations were performed and analysed using the GROMACS 4.0.3 package (van der Spoel et al 2005). The parm99 force field of AMBER with the ff99SB modification

was used to describe standard residues (Wang et al 2000; Hornak et al 2006). The parameters of phosphoserine were taken from (Craft and Legge 2005), while parameterization of trimethyllysine was done in-house following the same protocol used for phosphoserine (Craft and Legge 2005). Counter ions were added to neutralize the system. Solvent was explicitly represented with roughly 7,000 TIP3P water molecules (Jorgensen 1981) in a truncated octahedron box; a concentration of 150 mM of NaCl was added to mimic physiological conditions. The integration time step for the simulations was set to 2.0 fs and all chemical bonds involving Hydrogen atoms were restrained using the Lincs algorithm (Hess et al 1997; Hess 2008). Long-range interactions were treated using the Particle Mesh Ewald approach (Darden et al 1993; Essmann et al 1995) with a 1 nm direct space cut-off. Initially, the whole system was relaxed by performing 1,000 steps of energy minimization. Then, the system was gradually heated from 0 to 300 K during a 500 ps MD run imposing harmonic constraints of 1.0E + 04 kJ/mol nm^2 to the protein complex and a constant pressure of 1 atm. Final temperature and pressure was reached coupling the system to a Nose–Hoover thermostat (Nosé 1984; Hoover 1985) and a Parrinello–Rahman barostat (Parrinello and Rahman 1981; Nosé and Klein 1983), respectively. System's central of mass motion was linearly removed every 5 ps. Production runs were carried out for 60 ns for systems *i* and *ii*. For system *iii*, production runs were performed for 40 ns from which only the last 30 ns were used for analysis. Since phosphorylation at Ser10 is supposed to further perturb the systems, simulations of phosphorylated systems *iv* and *v* were extended up to 100 ns. System configurations were collected every 1 ps. The last 30 ns of each simulation were used for analysis.

All the dynamic properties reported were calculated using standard utility programs included in the Gromacs 4.0.3 release. Root mean square fluctuations (RMSF) and deviations (RMSD) were calculated on the Cα atoms of each residue.

Difference contacts map was calculated subtracting contact maps averaged over the last 30 ns of the simulations of systems *i* and *ii* as a normal matrix operation.

Electrostatic potentials were calculated using APBS (Baker et al. 2001). Molecular visualization and graphics were performed with VMD (Humphrey et al. 1996).

Distance cut-offs for hydrophobic contacts and salt bridges were set at 0.5 nm. Hydrogen bonds (Hbond) were considered to exist for acceptor–donor distances less than 0.3 nm and the angle acceptor–donor–hydrogen less than 30°.

Binding energies were calculated using an implicit solvation approach to take into account the instantaneous response of solvent dielectric. They were calculated as the difference between the energy of the complex and the sum of the energies of the isolated components along the MD runs. For this aim, we filtered out the trajectories of each complex and its components (water and counterions were striped out). Energies were evaluated within the Generalized Born Model framework as implemented in the sander module of Amber (Tsui and Case 2001). Notice that, strictly speaking, these values rather correspond to binding enthalpies. Furthermore, since energy values are calculated from an effective force field they should be taken as relative indicators of the strength of the interactions in each system and not as absolute values.

Results

Overall description

We used as template the NMR structure of the mouse HP1β Chromodomain in complex with the first 18 residues of H3 (Nielsen et al. 2002). To acquire a comparative overview of the isoform specific and post-transductional modification effects in the HP1–H3 complex we constructed five systems (see Sect. "Methods" for a more detailed description):

(i) The HP1β Chromodomain bound to the N-terminal 18-mer of H3 trimethylated at Lys9 (H3K9Me3).
(ii) The HP1γ Chromodomain bound to the N-terminal 18-mer of H3K9Me3.
(iii) The HP1β Chromodomain bound to the N-terminal 18-mer of H3K9Me3. In this system, residues *Glu16* and *18* at the Chromodomain were mutated to *Ala* and *Pro*, respectively. This was used as a control to test the involvement of acidic residues at the N-terminal of HP1β on the stability of the complex.
(iv) A complex of the human HP1β Chromodomain bound to the H3K9Me3 phosphorylated at Ser10 (H3K9Me3S10p).
(v) A complex of the human HP1γ Chromodomain bound to H3K9Me3S10p.

Note that aimed to ease the comparison between both isoforms, the numeration corresponding to the beta isoform is always used in the paper. The right residue numbering of the gamma isoform is shifted by 9 positions to the left (i.e. the first residue in our HP1γ model, which is called Glu15, corresponds actually to Glu24). Furthermore, in order to increase the comprehensiveness of the text, residues belonging to the Chromodomain receptor are hereafter reported in italics, while residues belonging to the H3 peptide are written using normal characters.

All the simulations were characterized by relatively large fluctuations, especially due to the presence of the

highly flexible segments at the N- and C-terminal regions of the H3 peptides. Root mean square deviation (RMSD) calculation for the Chromodomain of HP1β in complex with H3K9Me3 (system *i)* oscillated around 0.2 nm, which is well compatible with the 0.16 nm measured among the NMR family of structures. Measurement of RMSD for all the other systems studied gave slightly higher values ($\sim$0.3 nm). This could be expected as these systems were obtained as modifications from an experimental structure. Nevertheless, calculation of the cosine content of the first four eigenvectors, which account for more than 60% of the total motion in any of the systems, gave values below 0.4, suggesting a good convergence. The conformation of K9Me3 within the aromatic cage was maintained in all the systems studied. The overall agreement with the experimental data can be also inferred from the global match of the RMSF profiles for all the systems studied when compared with that obtained from the NMR derived structure (Fig. 2a).

In the following paragraphs we present a comparative view of the results obtained for the different HP1–H3 complexes. Conserved features already described in experimental structures are, in general, omitted.

HP1 isoform-specific interactions with the N-terminal of histone 3

System i: HP1β–H3K9Me3

The HP1β–H3K9Me3 complex reveals the minor modifications from the initial conformation (Fig. 1b). Secondary structure changes were not observed in the Chromodomain neither in the central residues Gln5 to Ser10 flanking K9Me3. This segment keeps a 5-residues long anti-parallel

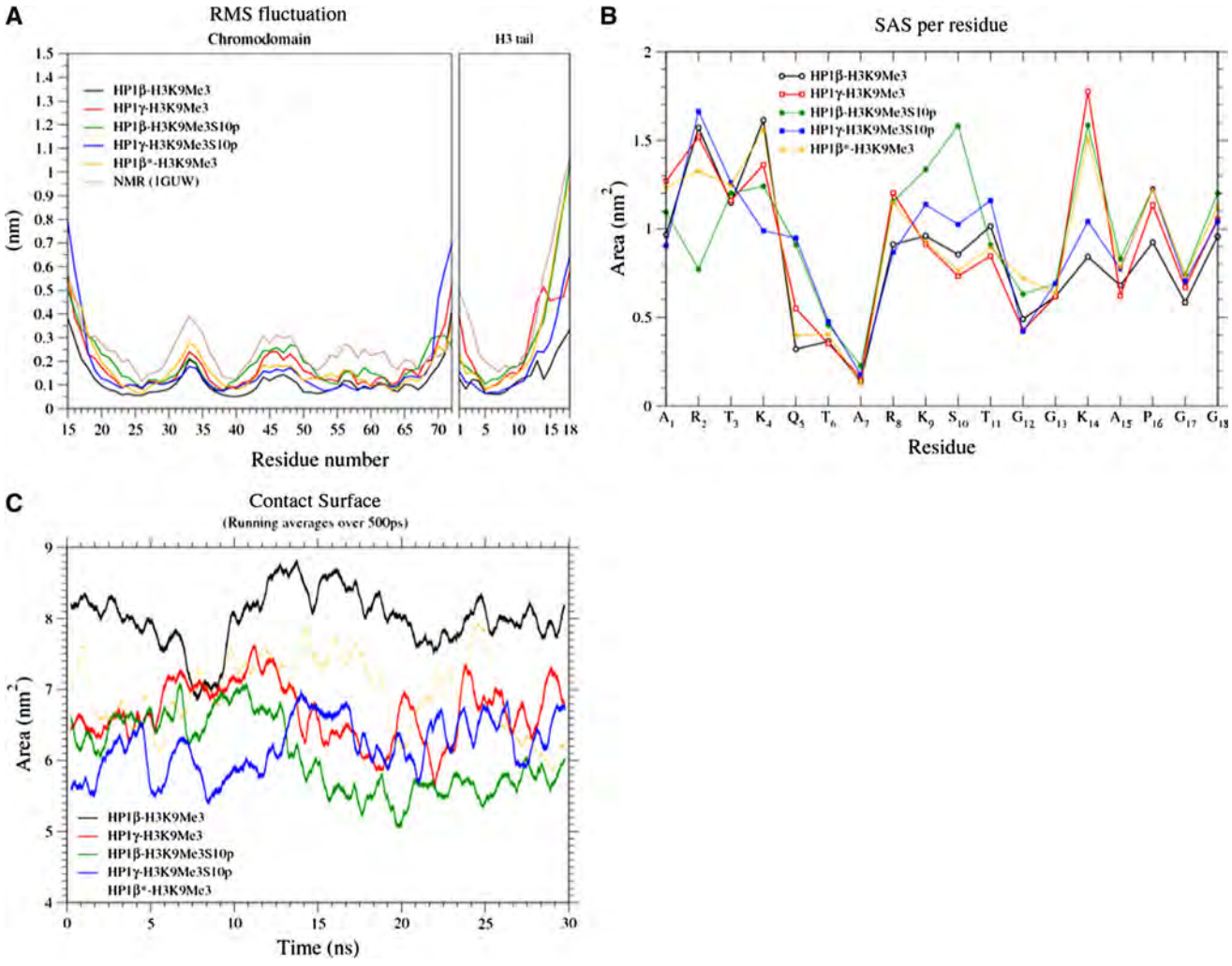

Fig. 2 Comparison of dynamical data extracted from the simulations. **a** RMSF calculated over the Cα atoms. Values for the Chromodomain and H3 peptide are presented on the left and right sides of the figure, respectively. Different systems are indicated by different colours. HP1β* corresponds to the double mutation, *Glu16Ala*, *Glu18Pro* introduced in HP1β (System *iii*). **b** Solvent accessible surface (SAS) area per residue of the ligand. **c** Instantaneous values of the protein–protein interface area for the MD trajectories

β-sheet conformation (Fig. 1a, b). This interaction is maintained by the formation of several hydrogen bonds between the backbone of the H3-tail and the Chromodomain (Table 1). Molecular dynamics simulation kept the same interactions found in the NMR structure. A global assessment of the fidelity of the MD trajectory when compared with the NMR data can be acquired by comparing the RMSF of the Cα atoms of each residue (Fig. 2a). We obtained a very good qualitative agreement with the peaks corresponding to the N- and C-termini and residues located in loops β1-β2 (residues *Lys33* to *Lys35*), β2-β3 (residues *Lys43* to *Asn50*) and β3-α1 (residues *Glu55* to *Cys60*, see secondary structure assignment in Fig. 1a). In agreement with the experimental data, the low mobility of residues 4 to 10 in the ligand peptide indicates the most stable interactions (Fig. 2a). Inspection of the MD trajectory also allows getting a relative measure of the strength of the interactions (Table 1) and account for the impairing effect of a series of point mutations reported for other Chromodomain/H3 interactions (Jacobs and Khorasanizadeh 2002). In particular, the stringent requirement for a Threonine at position 6 is justified by the simultaneous formation of an Hydrogen bond (Hbond) between its hydroxyl moiety with the carboxyl of *Glu20* and the hydrophobic interaction with the side chains of *Val22* of HP1 Chromodomain (Table 1 and Fig. 1c). The next residue, Ala7 is deeply buried in the protein–protein interface surrounded by the highly conserved *Val23*, *Leu40*, *Trp42* and *Leu58* (Table 1). This tight

Table 1 Comparison of main interactions and binding energies involved in HP1–H3 complexes

			HP1β				*HP1β*$_{E16A,E18P}$		*HP1γ*			
			K9Me3		*K9Me3S10p*		*K9Me3*		*K9Me3*		*K9Me3S10p*	
Binding energy estimation (kcal/mol)			−106.4		−94.7		−99.3		−94.6		−80.1	
Residue interactions												
	HP1	H3-tail	% Occ	$\tau_{1/2}$	% Occ	$\tau_{1/2}$	% Occ	$\tau_{1/2}$	% Occ	$\tau_{1/2}$	% Occ	$\tau_{1/2}$
Backbone												
HB	Val23:N	Gln5:O	63.1	3	58.6	3	44.3	2	61.5	3	64.6	3
	Asp59:O	Thr6:N	10.6	1	0.0	0	12.8	1	18.7	2	0.0	0
	Asp59:N	Thr6:O	65.3	4	0.0	0	70.2	4	50.3	4	0.0	0
	Tyr[Phe]21:O	Ala7:N	54.3	2	51.2	2	46.6	2	52.7	2	52.3	2
	Tyr[Phe]21:N	Ala7:O	61.3	3	58.9	3	55.7	2	67.4	3	69.0	4
	Asn57:O	Arg8:N	28.9	2	<0.1	1	53.0	3	22.0	2	0.0	0
	Glu19:O	K9Me3:N	0.6	1	61.3	3	49.5	4	3.3	1	0.0	0
Lateral chain												
HB	Asp62:N	Gln5:OE1	18.8	2	0.0	0	18.5	1	30.7	2	0.0	0
	Glu20:OE1(2)	Thr6:OG1	52.5	11	81.9	41	46.4	10	42.5	9	55.9	14
	Glu53:OE1(2)	Ser10:OG	9.4	11	–	–	89.4	22	78.3	43	–	–
	Asn57:ND2	Ser10:OG	10.7	2	0.0	0	0.2	1	1.5	2	0.0	0
SB	Glu20:CD	Arg8:CZ	70.0	14	10.0	5	32.3	12	61.8	9	0.4	2
	Glu19:CD		0.0	0	0.0	0	1.3	5	0.0	0	0.1	1
	Glu18:CD		43.5	3	35.2	15	–	–	–	–	–	–
	Glu17:CD		0.0	0	0.2	6	0.2	7	50.8	12	1.0	18
	Glu16:CD		52.6	24	5.2	4	–	–	–	–	–	–
	Glu15:CD		5.3	11	45.4	12	10.2	10	0.0	0	0.0	0
HC	Val22:CG1(2)	Thr6:CG2	99.0	105	75.5	6	98.9	109	95.5	39	99.9	810
	Val23:CG1(2)	Ala7:CB	99.9	832	99.6	325	99.9	856	79.3	8	97.3	46
	Trp42:CZ3		>99.9	5 ns	>99.9	3.3 ns	100	30 ns	98.9	99	99.9	881
	Trp42:CH2		>99.9	3.3 ns	>99.9	15 ns	>99.9	15 ns	99.0	109	>99.9	5 ns
	Leu40:CD1(2)		98.2	69	99.3	236	98.7	93	39.2	4	42.1	2
	Leu58:CD1(2)		78.3	24	72.1	15	98.8	99	60.8	16	82.9	8

Only interactions between HP1 and the β structured peptide of H3 are reported. Interactions involving K9Me3, which are conserved in all the simulations, are not included for the sake of brevity. The occurrence time (% Occ) was calculated as the percentage of time on which the interaction is observed over the last 30 ns of each trajectory. The average lifetime ($\tau_{1/2}$) of the interactions are reported in pico seconds (ps) unless ns (nano second) is indicated

HB hydrogen bonds, *HC* hydrophobic contacts, *SB* salt bridges

hydrophobic coordination combined with the reduced size of the hydrophobic cavity clearly explains the drastic reduction of an Ala7Met replacement (Jacobs and Khorasanizadeh 2002). Proceeding on the H3 sequence we found Arg8, the mutation of which into Alanine reduces Chromodomain binding by nearly two orders of magnitude (Jacobs and Khorasanizadeh 2002). In fact, this basic residue occupies a key position to interact with *Glu16, 18* and *20* (Table 1, Fig. 1c). Then, K9Me3, the trimethyl ammonium moiety of which remains tightly coordinated by the triple aromatic cage formed by *Tyr21*, *Trp42* and *Tyr45* along the whole simulation. Subsequently, Ser10, which is target of phosphorylation, establish in this isoform only transient electrostatic interactions with *Glu53* and *Asn57*, probably due to solvent competition. This furnishes a putative explanation for the mild affinity loss upon mutation into Alanine (Jacobs and Khorasanizadeh 2002). Another interesting residue is Lys14, which is target of acetylation (Yang 2004). In close similarity with Ser10, post-transductional modification at Lys14 has been proposed to mediate dissociation (Mateescu et al. 2004). A supposed explanation for this effect could be the neighbourhood of the amide moiety of Lys14 with the negative region generated by the six consecutive Glutamate residues at the N-terminal of the HP1β Chromodomain (Fig. 1c). However, these contacts are not stiffly maintained during the simulation, suggesting a rather unspecific interaction. It is, hence, possible that turning off the positive charge by acetylation may decrease the binding affinity by reducing the global coulombic attraction.

System ii: HP1γ–H3K9Me3

The simulation of the HP1γ isoform did not show dramatic changes in the global structure of the complex with respect to HP1β (Fig. 1b, d). The main structural distortion resides in the loss of the last helical turn, although it does not seem to have a direct effect in the intermolecular interactions. The anti-parallel β-sheet conformation in the core of the H3 peptide is kept within the binding site. Indeed, Hbond interactions involved in the central β-sheet remain essentially unchanged (Table 1). However, the lack of electrostatic interactions originated by the substitution of *Glu16* and *18* in HP1β for *Ala16* and *Pro18* in HP1γ generate some structural instability that translate in a slightly lower affinity. Calculation of the average binding energy during the trajectory showed a decrease of nearly 10% (Table 1). This is consistent with the behaviour of protein–protein interface area, which in HP1γ evolves to lower values with respect to the beta isoform (Fig. 2c). In line with this, the number of salt bridge interactions engaged by Arg8 is reduced (Table 1). Additionally, we observe an almost complete loss of interactions between the N-terminal region of the Chromodomain and Lys14. This situation can also be inferred from the increase in the solvent accessible surface of both residues as compared with the HP1β complex (Fig. 2b).

Absence of electrostatic stabilization translates in a higher flexibility (Fig. 2a). The largest differences in RMSF between both Chromodomain variants are observed in the region belonging to the loop β2-β3, which contains *Phe45*, one of the three aromatic residues involved in the cation-π interaction with K9Me3. Furthermore, a large rise in RMSF is observed on the C-terminal half of the peptide that moves freely without establishing any stiff interaction during the simulated time window (Fig. 2a).

A comparative overview of the dissimilarities concerning both isoforms can be acquired by inspection of the different contact maps averaged over the trajectory (Fig. 3). The diagonal elements of the symmetric map represent self-residue contacts, which are obviously always present and conserved. Off-diagonal elements are indicative of diverse inter-residue contacts along the trajectory of both isoforms. In this representation, green pixels (background) correspond to the zero in the scale and represents contacts conserved in both isoforms. Colours to the left in

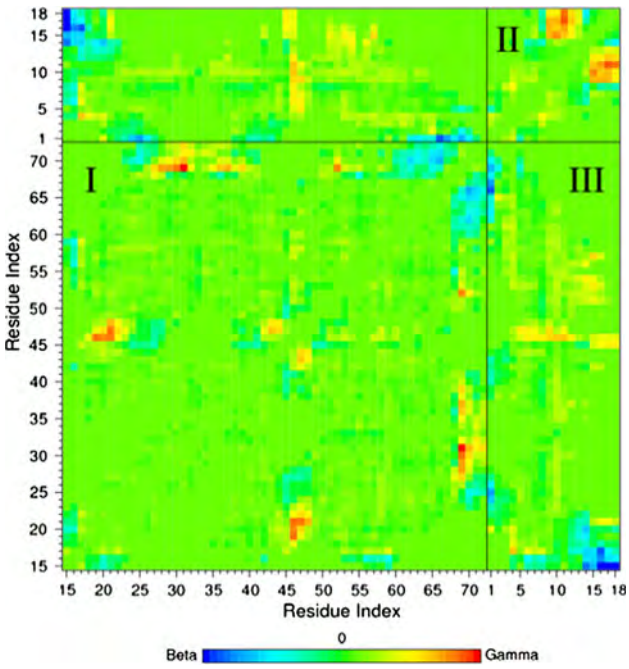

Fig. 3 Difference contact map. This map was obtained as the matrix difference between the contact map averaged over the last 30 ns of the trajectories of systems *i* and *ii*. The map is symmetric with respect to the positive diagonal. Each pixel corresponds to a single amino acid. *Green pixels* indicate contacts conserved in both simulations. The colour scale runs according to the light spectrum, where *blued pixels* are indicative of contacts observed only in the beta isoform and *yellow* or *reddish pixels* are observed only in the gamma isoform. Regions I, II and III contains intra-Chromodomain, intra-peptide and Chromodomain–peptide interactions, respectively

the scale (blued pixels) are indicative of inter-residue contacts present only in the HP1β. Yellow or reddish regions indicate contacts present solely in HP1γ (Fig. 3). The blued (reddish) the colour, the longer the occurrence of the inter-residue contacts during the simulation.

Inspection of the intra-Chromodomain interactions (region I in Fig. 3) indicate that only minor variations are observed going from one isoform to the other and these are due to sequence dissimilarities (see also Fig. 1a). The more evident contact variations within the Chromodomains regard the C-terminal of the protein (residues *Leu68* to *Lys72*). These regions interact preferentially with nearby amino acids *Glu62* to *Phe67* in HP1β while the same segment is in touch with residues *Leu27* to *Phe39* in HP1γ. This is due to the loss of secondary structure in that region (Fig. 1d) originating a more flexible C-terminal segment that is able to sample a wider conformational space. Furthermore, residues *Thr46* to *Ala48* interact with *Glu19* to *Val22* and *Trp42* to *Gly44* only in HP1γ.

The region of the map covering the intra-peptide interaction (region II in Fig. 3) reveals contacts between the middle-terminal segments of the ligand in HP1γ. These contacts are not present in HP1β due to the more stable interaction with Lys14, which anchors the C-terminal of the peptide.

Examination of the region III, corresponding to the Chromodomain–peptide, shows that due to the more rigid conformation of the peptide, the N- and C-termini of the HP1β Chromodomain engage stable interactions with the C- and N-termini of the ligand, respectively. Conversely, the disordered behaviour of the ligand in complex with HP1γ results in more spread contacting regions.

System iii: probing the role of isoform-specific electrostatic interactions

Comparative analysis of the simulations of systems *i* and *ii* points to a fundamental role of charged residues at the N-terminal of the Chromodomain. Aimed to test this hypothesis we constructed a model of an HP1β–H3K9Me3 complex in which we introduced the double mutation *Glu16Ala*, *Glu18Pro* (*System iii*). In this way we can mimic the loss of electrostatic stabilization in HP1γ within the HP1β context. In agreement with our hypothesis, this double mutant recovered the most characteristic features obtained for the HP1γ–H3K9Me3 complex. Not only RMSF shifted to higher values as observed when passing from the beta to the gamma isoform (Fig. 2a), but also the solvent accessible surfaces per residue (Fig. 2b) are strikingly alike to those of the peptide bound to HP1γ. The interaction energy and the complex interface area show intermediate values to those obtained in both isoforms (Table 1 and Fig. 2c). In fact, comparison of the global electrostatic/hydrophobic interactions reported in Table 1 indicates that this mutant is more similar to HP1γ. This strongly suggests that the isoform-specific differences observed between beta and gamma variants are largely attributable to the *Glu16Ala*, *Glu18Pro* mutations.

Effect of Ser10 phosphorylation

System iv: HP1β in complex with H3K9Me3S10p

Phosphorylation at Ser10 concomitantly with trimethylation at Lys9 in the H3 peptide clearly destabilized the structure of the complex with HP1β. This translates in a lower amount of Hbond interactions (Table 1). In particular, the anti-parallel β-sheet is shortened in H3 with respect to both non-phosphorylated forms, although there is a partial compensation due to the creation of a new Hbond between the backbone amide group of K9Me3 and *Glu20* (Table 1). Still, due to the augmented flexibility of the N- and C-termini of the peptide, the salt bridges between Arg8 and the acidic N-terminal of the Chromodomain are sensibly reduced as well as the rest of the Hbond interactions listed in Table 1. The reason for this behaviour is that upon post-transductional modification, S10p increases significantly its solvent accessible surface. This slightly pulls out K9Me3, which increments its solvent accessible surface (Fig. 2b). Additionally, the separation of S10p drives the disruption of the electrostatic interactions established by Lys14 raising the solvent exposure of this residue as found in systems *ii* and *iii* (Fig. 2b).

Calculation of the protein–protein interface indicates a significant reduction upon phosphorylation (Fig. 2c), resulting even lower than the interface area measured for HP1γ. Nevertheless, we were not able to identify a clear indicator of complex dissociation within the time scale explored. Indeed, the hydrophobic interactions established by Ala7, which is deeply buried in the protein–protein interface, remain essentially unchanged (Table 1, Fig. 2b). To further investigate this issue, we calculated the back projection of the first 8 eigenvectors on the real space trajectory, which account for more than 70% of the total motion. However, no significant component of the movement was found onto the line determined by the centres of mass of both ligands.

System v: HP1γ in complex with H3K9Me3S10p

Introduction of phosphorylation at Ser10 in HP1γ has similar effects to those described for the beta isoform. Analogously, binding energy decreased nearly a 10% with respect to the unphosphorylated case (Table 1). The anti-parallel β-strand of H3 also shortens by two residues and, in general, all the interactions listed in Table 1 show a similar

variation as observed for the beta isoform. Also in this case no sign of dissociation was evident from the simulation. The main difference with the latter case resides in the interaction of Lys14, which is involved in fleeting electrostatic interactions with *Glu53* and *Asn57*. Similar interactions established between these residues and Ser10 were observed in systems *i* and *ii* (unphosphorylated forms). These interactions result in a smaller decrease in the protein–protein interface area when compared with system *iv* (Fig. 2c).

Discussion and conclusions

Transcription of the integrated HIV-1 provirus is ruled by chromatin organization, host cell transcription factors and chromatin modifying complexes that may promote the formation of a latent viral reservoir. The latent HIV-1 proviral 5′ LTR is organized into a defined structure composed by two positioned nucleosomes flanking the enhancer region. Besides the non-acetylated state of these LTR-associated nucleosomes, they further suffer H3 trimethylation at Lys9, which cause transcriptional silencing upon the recruitment of HP1 (Sadowski et al. 2008). It has been pointed that the gamma isoform of HP1 is a main determinant of the chromatin-mediated HIV-1 transcriptional silencing and post-integration latency (du Chéné et al. 2007). More recent evidence has suggested a kind of switching mechanism in which HP1β is replaced by the HP1γ isoform (Mateescu et al. 2008). In this context, the structural characterization of the isoform-specific interactions that define the binding preference for the trimethylated H3 tail is very important for a better understanding of the processes that rule the epigenetic control and for the rational design of small molecules able to selectively disrupt such protein–protein interactions. Aimed to provide structural insights into these interactions we have presented here a series of molecular simulations of the HP1 Chromodomain (isoforms beta and gamma) in complex with the N-terminal tail of H3 performed under homogeneous conditions. We also investigated the role of H3 phosphorylation at Ser10 since this modification has been proposed to mediate HP1–H3 dissociation (Fischle et al. 2005; Hirota et al. 2005; Johansen and Johansen 2006).

Sequence alignment of the three human HP1 isoforms indicates overall high identity conservation, especially in the structured domains (Fig. 1a). Although binding is mainly determined by the cation-π interaction, it is expected that isoform-specific interactions may modulate the molecular recognition. In qualitative agreement with experimental reports, we found better stabilization energy for the beta isoform adduct (Fischle et al. 2005). Our results underline the relevance of non-conserved residues at the N-terminal of the HP1 Chromodomain for the H3 binding although they are not expected to have any structural consequence for the structure of the isolated Chromodomains. Notably, these residues interact with Lys14 at H3, which is target of acetylation. This electrostatic interaction results very important for the stability of the bound peptide, which remains in a more extended and stable conformation in the complex with HP1β, while it results more flexible in the HP1γ adduct. This interaction seems to account for most of the isoform-specific effects since a very similar binding pattern is retrieved by introducing the two *Glu16Ala*, *Glu18Pro* mutations in the HP1β Chromodomain. The specificity of these effects is highlighted by the fact that these modifications are present only in the gamma isoform, while Glutamate residues are conserved in the alpha and beta variants (Fig. 1a).

Introduction of phosphorylation at Ser10 translate into a putatively less stable interaction in both isoform complexes. This can be seen from a higher mobility of the peptide segments at N- and C-terminal of K9Me3, which remain stably bound to the aromatic cage. Although we observed a reduction in the Hbond interactions that anchor the H3 β-strand to the binding site, a reduced binding energy, a slight increase in the solvent accessible surface of the binding peptide and a reduction of the protein–protein interface area, the ligands remained bound without an evident tendency to dissociation. In line with this observation, several experimental studies indicate that S10p may not be enough for dissociation to happen (Mateescu et al. 2004). It has also been suggested that concomitant acetylation at Lys14 is needed to detach H3 from HP1 (Mateescu et al. 2004). If this were the case, our results would suggest that the lack of interactions between Lys14 and the acidic residues at the N-terminal could be part of the release mechanism. Furthermore, it could be conjectured that this modification might be less effective in the context of the gamma isoform, where Ser10 phosphorylation could be better tolerated. This can be inferred from the electrostatic potential to which the binding peptides are exposed by the receptor. As illustrated in Fig. 4, the electrostatic potential generated by the HP1β isoform is more negative, especially in the region surrounding Ser10.

Of course, we have to keep in mind some intrinsic limitations of the theoretical methods, such as the limited sampling time, absence of polarization effects, rough description of cation-π interaction, etc. We also have to underline that a strong bias is imposed in the simulations of phosphorylated systems by assuming that the doubly modified peptides are bound to the Chromodomain. Another, perhaps the more important, shortcut regards the suboptimal reproduction of interactions arising from the lack of biological environment whose effects are impossible to estimate. Nevertheless, state of the art simulations as those presented in this contribution performed under homogeneous conditions may help to pin

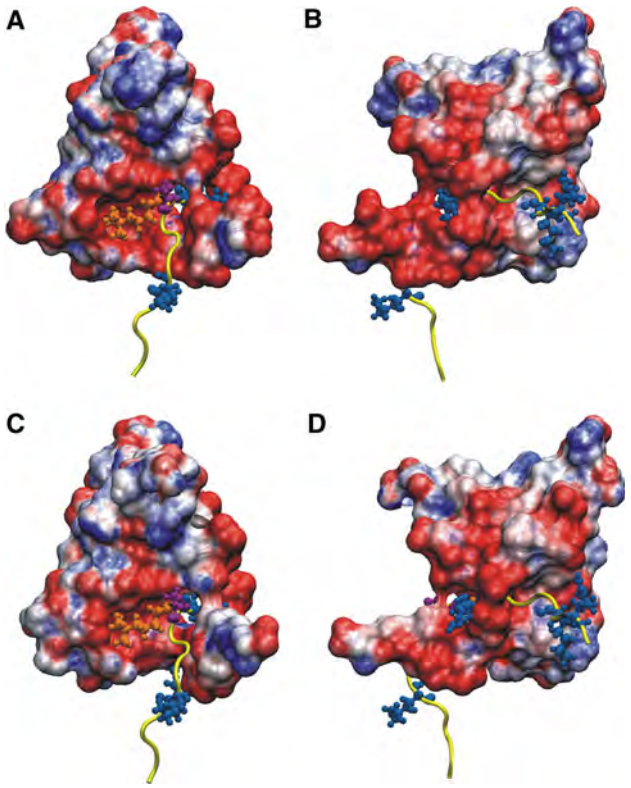

Fig. 4 Electrostatic potential mapped on the HP1 solvent accessible surfaces (Connolly type surface, Varshney et al 1994). The electrostatic potential was calculated only for the Chromodomain receptors previous to each MD to allow for a better comparison. Molecular representations of the ligand are included in the figure but they were not used in the calculation. *Red*, *white* and *blue* regions correspond to negative, neutral and positive potential, respectively. K9Me3 residue is coloured in *orange*, while Ser10 and all the basic residues present in the H3 peptide (Arg2, 8 and Lys4, 14) are coloured in *purple* and *blue*, respectively. **a** and **b** Electrostatic potential for HP1β rotated 180° around the vertical axis. **c** and **d** same as **a** and **b** for HP1γ

down isoform-specific interactions that define the binding preference for a given target. In particular, the absence of acidic residues at the N-terminal segment of HP1γ may be exploited as a selectivity determinant for the rational design of small molecules able to selectively disrupt these protein–protein interactions.

Acknowledgments 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 n° DCI – ALA / 2007 / 19.040 between Uruguay and the European Commission) and Grant FCE_60-2007. M. R. M. is a beneficiary of the National Fellowship System of ANII.

References

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041

Ball LJ, Murzina NV, Broadhurst RW, Raine AR, Archer SJ, Stott FJ, Murzin AG, Singh PB, Domaille PJ, Laue ED (1997) Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. EMBO J 16:2473–2481

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410:120–124

Clavel F, Hance AJ (2004) HIV drug resistance. N Engl J Med 350:1023–1035

Craft JW, Legge GB (2005) An AMBER/DYANA/MOLMOL phosphorylated amino acid library set and incorporation into NMR structure calculations. J Biomol NMR 33:15–24

Darden T, York D, Pedersen L (1993) Particle Mesh Ewald: an Nlog(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092

du Chéné I, Basyuk E, Lin YL, Triboulet R, Knezevich A, Chable-Bessia C, Mettling C, Baillat V, Reynes J, Corbeau P, Bertrand E, Marcello A, Emiliani S, Kiernan R, Benkirane M (2007) Suv39H1 and HP1gamma are responsible for chromatin-mediated HIV-1 transcriptional silencing and post-integration latency. EMBO J 26:424–435

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593

Fischle W, Tseng BS, Dormann HL, Ueberheide BM, Garcia BA, Shabanowitz J, Hunt DF, Funabiki H, Allis CD (2005) Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. Nature 438:1116–1122

Grewal SI, Moazed D (2003) Heterochromatin and epigenetic control of gene expression. Science 301:798–802

Hess B (2008) P-LINCS: a parallel linear constraint solver for molecular simulation. J Chem Theory Comput 4:116–122

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472

Hirota T, Lipp JJ, Toh BH, Peters JM (2005) Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. Nature 438:1176–1180

Hoover WG (1985) Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 31:1695–1697

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65:712–725

Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33–38

Jacobs SA, Khorasanizadeh S (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science 295:2080–2083

Jacobs SA, Taverna SD, Zhang Y, Briggs SD, Li J, Eissenberg JC, Allis CD, Khorasanizadeh S (2001) Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. EMBO J 20:5232–5241

Johansen KM, Johansen J (2006) Regulation of chromatin structure by histone H3S10 phosphorylation. Chromosome Res 14:393–404

Jorgensen WL (1981) Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. J Am Chem Soc 103:335–340

Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410:116–120

Lehrman G, Hogue IB, Palmer S, Jennings C, Spina CA, Wiegand A, Landay AL, Coombs RW, Richman DD, Mellors JW, Coffin JM, Bosch RJ, Margolis DM (2005) Depletion of latent HIV-1 infection in vivo: a proof-of-concept study. Lancet 366:549–555

Maison C, Almouzni G (2004) HP1 and the dynamics of heterochromatin maintenance. Nat Rev Mol Cell Biol 5:296–304

Marban C, Suzanne S, Dequiedt F, de Walque S, Redel L, Van Lint C, Aunis D, Rohr O (2007) Recruitment of chromatin-modifying enzymes by CTIP2 promotes HIV-1 transcriptional silencing. EMBO J 26:412–423

Martinez-Cajas JL, Wainberg MA (2008) Antiretroviral therapy: optimal sequencing of therapy to avoid resistance. Drugs 68:43–72

Mateescu B, England P, Halgand F, Yaniv M, Muchardt C (2004) Tethering of HP1 proteins to chromatin is relieved by phosphoacetylation of histone H3. EMBO Rep 5:490–496

Mateescu B, Bourachot B, Rachez C, Ogryzko V, Muchardt C (2008) Regulation of an inducible promoter by an HP1beta-HP1gamma switch. EMBO Rep 9:267–272

McKinnon JE, Mellors JW, Swindells S (2009) Simplification strategies to reduce antiretroviral drug exposure: progress and prospects. Antivir Ther 14:1–12

Minc E, Courvalin JC, Buendia B (2000) HP1gamma associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. Cytogenet Cell Genet 90:279–284

Nielsen AL, Oulad-Abdelghani M, Ortiz JA, Remboutsika E, Chambon P, Losson R (2001) Heterochromatin formation in mammalian cells: interaction between histones and HP1 proteins. Mol Cell 7:729–739

Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature 416:103–107

Nosé S (1984) A molecular dynamics method for simulations in the canonical ensemble. Mol Phys 52:255–268

Nosé S, Klein ML (1983) Constant pressure molecular dynamics for molecular systems. Mol Phys 50:1055–1076

Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 52:7182–7190

Pierson T, McArthur J, Siliciano RF (2000) Reservoirs for HIV-1: mechanisms for viral persistence in the presence of antiviral immune responses and antiretroviral therapy. Annu Rev Immunol 18:665–708

Sadowski I, Lourenco P, Malcolm T (2008) Factors controlling chromatin organization and nucleosome positioning for establishment and maintenance of HIV latency. Curr HIV Res 6:286–295

Tsui V, Case DA (2001) Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopolymers 56:275–291

van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible and free. J Comp Chem 26:1701–1718

Varshney A, Brooks FP, Wright WV (1994) Computing smooth molecular surfaces. IEEE Comput Graph Appl 14:19–25

Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 21:1049–1074

Yang XJ (2004) Lysine acetylation and the bromodomain: a new partnership for signaling. Bioessays 26:1076–1087

Ylisastigui L, Archin NM, Lehrman G, Bosch RJ, Margolis DM (2004) Coaxing HIV-1 from resting CD4 T cells: histone deacetylase inhibition allows latent viral expression. AIDS 18:1101–1108