

UNIVERSIDAD DE OVIEDO

Departamento de Química Física y Analítica

Área de Química Física

**Avances Metodológicos en el
Cálculo de la Entropía Conformacional
y la Energía de Biomoléculas
y su Aplicación a Modelos de Colágeno**

Tesis Doctoral

por

Ernesto Suárez Álvarez

2011

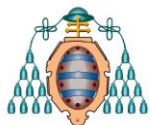


RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Avances Metodológicos en el Cálculo de la Entropía Conformacional y la Energía de Biomoléculas y su Aplicación a Modelos de Colágeno	Inglés: Methodological Improvements for Computing the Conformational Entropy and Energy of Biomolecules and their Application to Collagen Models
2.- Autor	
Nombre: Ernesto Suárez Álvarez	DNI/Pasaporte/NIE:
Programa de Doctorado cursado: Química Teórica y Modelización Computacional	
Órgano responsable: Departamento de Química Física y Analítica (Programa interuniversitario)	

RESUMEN (en español)

En esta Tesis se persigue introducir algunas mejoras en el estudio teórico de biomoléculas mediante dos estrategias. En primer lugar, se propone un método para calcular la entropía conformacional de una molécula a partir de simulaciones de dinámica molecular, asumiendo que la entropía total intra-molecular (excluyendo rotación y traslación) es separable en sendas componentes vibracional y conformacional. La entropía vibracional se calcula utilizando la clásica aproximación harmónica. Para evaluar la componente conformacional, y partiendo de la conocida expansión en términos de funciones de información mutua, se han desarrollado una serie de nuevos métodos que son capaces de capturar los efectos de la correlación a órdenes elevados dentro de un *cutoff* dado. Así, para sistemas de gran tamaño, se propone utilizar el denominado *correlation-corrected multibody local approximation*, que incorpora únicamente la correlación verdadera o genuina presente en una determinada simulación. En segundo lugar, en esta Tesis se persigue el refinamiento de las energías de biomoléculas obtenidas generalmente mediante la utilización de potenciales de la mecánica molecular. Para ello se propone una aproximación termoquímica que estima la energía mecanocuántica de toda la biomolécula a partir de las energías de sus fragmentos y que puede combinarse fácilmente con cualquier modelo mecanocuántico de disolvente continuo o con métodos híbridos. Los dos avances metodológicos propuestos se aplican principalmente durante el estudio de modelos de la triple hélice de colágeno, que es la proteína más abundante en el medio extracelular de los mamíferos. También se han analizado otros sistemas, como alcanos en fase gas o polipéptidos en forma libre o unidos a enzimas metaloproteinasas, para los que es posible hacer cálculos de entropía más exigentes que facilitan el análisis de los resultados y/o la comparación con datos experimentales.



RESUMEN (en Inglés)

The main goal of this Thesis is to improve the theoretical studies of biomolecules following two strategies. First, we propose a new method to compute conformational entropies of single molecules from molecular dynamics simulations assuming that the total intra-molecular entropy (excluding rotation and translation) can be separated into vibrational and conformational contributions. The vibrational part of the entropy can be estimated by the classical harmonic approach. By using the so-called Mutual Information Expansion as the starting point for computing the conformational entropy part, we have developed a series of new entropy methods that capture high order correlation effects within a predefined cutoff. More particularly, we propose to employ for large systems the correlation-corrected multibody local approximation that recovers only true correlation effects from the available amount of sampling. Secondly, we seek to refine the energy estimations for biomolecules that are generally computed using classical potentials. On the basis of thermochemical arguments, we propose a new fragment energy method which can be useful for, and readily applicable to biomolecules, using either quantum mechanical or hybrid quantum mechanical/molecular mechanics methods. The two methodological approaches have been mainly used in the study of triple helical peptides that mimic collagen structure, which is the most abundant protein in the extracellular environment of mammals. In addition, we have also studied other systems such as alkanes in the gas phase and polypeptides both in their unbound and enzyme-bound states, for which more demanding entropy calculations can be carried out, allowing thus deeper analyses of the results and/or better comparisons with experimental data.

SR. DIRECTOR DE DEPARTAMENTO DE _____ /
SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN _____

*Gracias a todos los que han contribuido a hacer
de esta Tesis un trabajo más llevadero*

RESUMEN

En esta Tesis se persigue introducir algunas mejoras en el estudio teórico de biomoléculas mediante dos estrategias. En primer lugar, se propone un método para calcular la entropía conformacional de una molécula a partir de simulaciones de dinámica molecular, asumiendo que la entropía total intra-molecular (excluyendo rotación y traslación) es separable en sendas componentes vibracional y conformacional. La entropía vibracional se calcula utilizando la clásica aproximación armónica. Para evaluar la componente conformacional, y partiendo de la conocida expansión en términos de funciones de información mutua, se han desarrollado una serie de nuevos métodos que son capaces de capturar los efectos de la correlación a órdenes elevados dentro de un *cutoff* dado. Así, para sistemas de gran tamaño, se propone utilizar el denominado *correlation-corrected multibody local approximation*, que incorpora únicamente la correlación verdadera o genuina presente en una determinada simulación. En segundo lugar, en esta Tesis se persigue el refinamiento de las energías de biomoléculas obtenidas generalmente mediante la utilización de potenciales de la mecánica molecular. Para ello se propone una aproximación termoquímica que estima la energía mecanocuántica de toda la biomolécula a partir de las energías de sus fragmentos y que puede combinarse fácilmente con cualquier modelo mecanocuántico de disolvente continuo o con métodos híbridos. Los dos avances metodológicos propuestos se aplican principalmente durante el estudio de modelos de la triple hélice de colágeno, que es la proteína más abundante en el medio extracelular de los mamíferos. También se han analizado otros sistemas, como alcanos en fase gas o polipéptidos en forma libre o unidos a enzimas metaloproteinasas, para los que es posible hacer cálculos de entropía más exigentes que facilitan el análisis de los resultados y/o la comparación con datos experimentales.

ABSTRACT

The main goal of this Thesis is to improve the theoretical studies of biomolecules following two strategies. First, we propose a new method to compute conformational entropies of single molecules from molecular dynamics simulations assuming that the total intra-molecular entropy (excluding rotation and translation) can be separated into vibrational and conformational contributions. The vibrational part of the entropy can be estimated by the classical harmonic approach. By using the so-called Mutual Information Expansion as the starting point for computing the conformational entropy part, we have developed a series of new entropy methods that capture high order correlation effects within a predefined cutoff. More particularly, we propose to employ for large systems the correlation-corrected multibody local approximation that recovers only true correlation effects from the available amount of sampling. Secondly, we seek to refine the energy estimations for biomolecules that are generally computed using classical potentials. On the basis of thermochemical arguments, we propose a new fragment energy method which can be useful for, and readily applicable to biomolecules, using either quantum mechanical or hybrid quantum mechanical/molecular mechanics methods. The two methodological approaches have been mainly used in the study of triple helical peptides that mimic collagen structure, which is the most abundant protein in the extracellular environment of mammals. In addition, we have also studied other systems such as alkanes in the gas phase and polypeptides both in their unbound and enzyme-bound states, for which more demanding entropy calculations can be carried out, allowing thus deeper analyses of the results and/or better comparisons with experimental data.

A mi familia

Índice General

Capítulo I:

Introducción 3

1.1	Cálculos de Entropía en Biomoléculas	6
1.1.1	Análisis Cuasi-harmónico: Primera Aproximación	7
1.1.2	Método de Schlitter	8
1.1.3	Análisis Cuasi-harmónico: Segunda Aproximación	10
1.1.4	Método NN (Nearest-Neighbor)	12
1.1.5	Método MIE (Mutual Information Expansion)	14
1.1.6	Método MCSA (Minimally Coupled Subspace Approach)	15
1.1.7	Métodos HS (Hypothetical Scanning)	16
1.1.8	Nuestra Propuesta para el Cálculo de Entropías Absolutas	17
1.2	Cálculos de Energía en Biomoléculas basados en Fragmentos	18
1.2.1	Método MBE (Multi-Body Expansion)	18
1.2.2	Método KEM (Kernel Energy Method)	21
1.2.3	Métodos Corregidos con un Entorno Electroestático	21
1.2.4	Método MFCC (Molecular Fractionation with Conjugate Caps) y Método MTA (Molecular Tailoring Approach)	22
1.2.5	Método de Fragmentación Sistemática	23
1.2.6	Observaciones Generales sobre los distintos Métodos	24
1.2.7	Nuestra Propuesta para el Cálculo de Energías basadas en Fragmentos	26
1.3	La Triple Hélice de Colágeno	27
1.3.1	Relevancia y función del Colágeno	27
1.3.2	Estructura de la Molécula de Colágeno	28
1.3.3	Modelos de Colágeno y Estabilidad de la Triple Hélice	30
1.3.4	Estudios Teóricos previos sobre Modelos de Colágeno	36
1.4	Cálculos de Energía y Entropía en Modelos de Colágeno: Desafíos Metodológicos	38

Objetivos 41

Capítulo II:

Discusión de Resultados y Publicaciones 43

2.1	Cálculos de Entropía en Biomoléculas a partir de Dinámica Molecular	44
2.1.1	Compendio de Publicaciones	51
2.1.1.1	<i>Entropic Control of the Relative Stability of Triple-helical Collagen Peptide Models</i>	53
2.1.1.2	<i>Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and QM/MM Calculations</i>	75
2.1.1.3	<i>Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations</i>	97

2.1.1.4 <i>Distinguishability in Entropy Calculations: Chemical Reactions, Conformational and Residual Entropies</i>	129
2.1.1.5 <i>Multibody Local Approximation for Conformational Entropy Calculations on Biomolecules</i>	139
2.1.1.6 <i>CENCALC: A New Program for Conformational Entropy Calculation of Macromolecules from Molecular Dynamics Simulation</i>	189
2.1.2 Otros Cálculos de Entropía Conformacional en Modelos de Colágeno	251
2.1.2.1 Modelos POG10 y T3-785	252
2.1.2.2 Modelo THP-1	255
2.1.2.3 Modelo fTHP-5	261
2.1.2.4 Observaciones Generales	265
2.2 Cálculos de Energía de Biomoléculas a partir de Fragmentos	267
2.2.1 Compendio de Publicaciones	271
2.2.1.1 <i>Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide</i>	273
2.2.1.2 <i>Thermochemical Fragment Energy Method for Quantum Mechanical Calculations on Biomolecules</i>	291

Conclusiones..... 303

Informe sobre el Factor de Impacto de las Publicaciones Presentadas307

Bibliografía 309

Capítulo I

Introducción

El estudio teórico de biomoléculas no es sólo un tema de un extraordinario interés,*(1-11)* sino también de una extraordinaria complejidad.*(12-20)* El relativo gran tamaño que suelen tener estos sistemas dificulta tanto la evaluación de magnitudes asignables a un solo punto del espacio de fases, como aquellas, encabezadas por la entropía, que caracterizan todo su volumen. El uso de potenciales de mecánica molecular, en los que se utilizan expresiones simples que son parametrizadas para reproducir la superficie de energía potencial del sistema, simplifica considerablemente el cálculo de magnitudes como la energía, su gradiente, los momentos lineales, etc. En cambio, la evaluación de la entropía y cualquier otra magnitud dependiente de ella, sigue siendo un reto dentro de la Química Computacional.

Al ser sistemas que se encuentran en un medio natural acuoso, el modelado del disolvente es otro problema añadido. En general, el número de moléculas de agua necesarias para simular adecuadamente este entorno suele ser elevado. Pero además, el disolvente tiene un espacio configuracional muy amplio que es prácticamente imposible de explorar adecuadamente, sobretodo, si el objetivo es evaluar la entropía del disolvente o su variación en procesos de solvatación. La principal consecuencia de todo esto, es que nos vemos obligados a utilizar modelos muy simples para estimar energías libres de solvatación. En cualquier caso, la evaluación de esta energía está fuera del alcance de la presente Tesis.

En esta Tesis se persigue introducir mejoras en el estudio teórico de biomoléculas mediante dos estrategias complementarias. En primer lugar se propone un método para calcular la entropía conformacional, la cual ha sido frecuentemente ignorada por

muchos métodos aproximados que, por ejemplo, estiman la estabilidad relativa de sistemas biológicos haciendo uso de entropías obtenidas a partir de cálculos de modos normales.(21-23) Pero la entropía conformacional es, justamente, la parte de la entropía que se pierde al utilizar la clásica aproximación armónica sobre los mínimos de la superficie de energía potencial. La partición de la entropía intramolecular en vibracional y conformacional simplifica enormemente, como veremos, el reto de la estimación de la entropía total. Aun así, la resolución del problema no es precisamente simple debido a la abrumadora dimensionalidad del espacio conformacional y a la inexistencia de estimadores no sesgados para la entropía.

En segundo lugar, nuestro objetivo es evaluar la posibilidad de refinar las energías de estos sistemas, obtenidas generalmente mediante la utilización de potenciales clásicos. Se sigue un método que estima la energía electrónica de grandes moléculas a partir de las energías de sus fragmentos. La naturaleza de este método nos permite, además, combinarlo trivialmente con casi cualquier tipo de modelo de disolvente continuo. Como veremos, la realización de re-cálculos sobre un conjunto relativamente pequeño de estructuras permitirá obtener valores promedio de energías mecanocuánticas con una razonable incertidumbre estadística.

Los dos avances metodológicos propuestos en esta Tesis se aplicarán principalmente durante el estudio de la triple hélice de colágeno. Para ello se han seleccionado varios modelos con distinta secuencia de aminoácidos, que se simularán mediante dinámica molecular. A partir de dichas simulaciones, se realizarán distintos análisis energéticos y entrópicos de las configuraciones en triple hélice o de las formas desnaturalizadas de las mismas. Veremos cómo estos modelos de colágeno son unos sistemas ideales en los que estudiar los cambios en la entropía conformacional debido a que en su formación se pasa de un estado *random-coil*, relativamente flexible y desestructurado, a una estructura esencialmente rígida como es la triple hélice. También se ha contemplado la utilización de los cálculos de entropía conformacional en sistemas bastante más pequeños que los modelos de colágeno seleccionados, como pequeños péptidos o alcanos, con el objetivo de validar y analizar el comportamiento de los nuevos métodos desarrollados.

Seguidamente haremos una revisión de los diferentes protocolos que se pueden encontrar en la literatura, tanto para el cálculo de entropías en general como para el cálculo de energías a partir de fragmentos. A continuación se presentarán las principales características (relevancia, estructura, estabilidad, etc.) y resultados teóricos previos obtenidos para la molécula de colágeno. Finalmente, enunciaremos los objetivos del presente trabajo en relación a los tres aspectos antes considerados.

1.1 Cálculos de Entropía en Biomoléculas

La entropía es una propiedad fundamental involucrada en todo evento natural, siendo muchas veces determinante en procesos biológicos elementales como el plegamiento de proteínas, los procesos de *binding* y la catálisis enzimática.(14, 24-28) Sin embargo, desde que la famosa ecuación de Boltzmann fue escrita por primera vez, la estimación por métodos teóricos de esta magnitud tan “común”, ha sido siempre un problema extremadamente difícil de tratar. Dado el tamaño de los sistemas que somos capaces de estudiar hoy en día, la continua búsqueda de métodos para el cálculo de la entropía mejores y más eficientes no parece tener fin. (12-13, 29-34)

Los métodos de estimación de la entropía que consideramos más relevantes pueden dividirse en las siguientes tres categorías:

- a) Métodos perturbativos que son normalmente utilizados para evaluar las diferencias de energía libre entre dos estados. Aquí se incluyen métodos como la integración termodinámica TI (*Thermodynamic Integration*) y la perturbación de energía libre FEP (*Free Energy Perturbation*). (35-36)
- b) Métodos paramétricos que se basan en una definición *a priori* de una forma funcional paramétrica para la función de densidad de probabilidad (FDP) que rige a las variables del sistema. Esta categoría está formada por métodos basados en el análisis cuasi-harmónico (QH). (12, 29-30, 37-38)
- c) Métodos no paramétricos donde no se hace ninguna consideración acerca de la FDP. Este es el caso del método MIE(32, 39) (*Mutual Information Expansion*) y de los estimadores NN (Nearest-Neighbor). (40-41)

Los métodos perturbativos pueden ser muy útiles para testar otras aproximaciones en sistemas de pequeño tamaño para los que conozcamos la energía libre asociada a algún estado de referencia adecuado. Por ejemplo, se puede plantear una transformación desde nuestro sistema concreto a un estado de referencia donde cada partícula oscila de manera independiente en un potencial armónico (ver Figura 3 en la referencia (42)), ya que al menos para esta referencia, podemos calcular la energía libre analíticamente tal y

como fue propuesto originalmente por Tyka y *col.*(43) De esta forma, si calculamos la variación de energía libre entre ambos sistemas por un método perturbativo, podríamos entonces dar un valor *absoluto* de energía libre para ambos extremos (y no sólo para la referencia).

En el caso de moléculas de gran tamaño, sin embargo, estos métodos sólo son viables si la varianza estructural entre los dos estados es muy pequeña. De lo contrario el coste computacional sería inasumible. En consecuencia, a continuación nos centraremos principalmente en describir las categorías (b) y (c) donde la FDP es estimada estadísticamente a partir de un muestreo del espacio configuracional mediante trayectorias de dinámica molecular o métodos de Monte-Carlo.

1.1.1 Análisis Cuasi-harmónico: Primera Aproximación

El aquí denominado análisis cuasi-harmónico fue propuesto por primera vez por Karplus y Kushick. Estos autores mostraron que la diferencia de entropía *configuracional* (no cinética) entre dos estados de una misma molécula puede ser estimada a partir de sus respectivas matrices de covarianza.(29) La idea central era considerar la densidad de probabilidad de las variables del sistema $P(\mathbf{q})$ como una función normal multivariable del tipo:

$$P(\mathbf{q}) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\sigma})^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{q} - \langle \mathbf{q} \rangle)^T \boldsymbol{\sigma}^{-1} (\mathbf{q} - \langle \mathbf{q} \rangle)\right] \quad (1.1)$$

donde $\mathbf{q} = (q_1, q_2, \dots, q_n)$ son las n coordenadas del sistema, $\langle \mathbf{q} \rangle$ contiene sus valores medios muestrales y $\boldsymbol{\sigma}$ es la matriz de covarianzas siendo $\det(\boldsymbol{\sigma})$ su determinante. La entropía es proporcional al valor esperado $E(\cdot)$ ¹ del logaritmo de la función de probabilidad cambiado de signo según la expresión

$$S = k_B E(-\ln P(\mathbf{q})), \quad (1.2)$$

¹ Hemos utilizado el símbolo $E(\cdot)$ y no el de $\langle \cdot \rangle$ porque este último lo reservaremos para medias muestrales.

siendo k_B la constante de Boltzmann. Entonces, para el caso continuo, la entropía configuracional sería

$$S_{Config} = -k_B \int_{\Omega} P(\mathbf{q}) \ln P(\mathbf{q}) d\mathbf{q}, \quad (1.3)$$

donde Ω es el espacio configuracional d -dimensional (\mathbb{R}^d en coordenadas cartesianas). Resolviendo la integral anterior utilizando la ecuación (1.1) obtenemos como resultado final:

$$S_{Config} = \frac{1}{2} k_B \left[n + \ln \left((2\pi)^n \det(\boldsymbol{\sigma}) \right) \right]. \quad (1.4)$$

Desde el principio, sin embargo, la implementación del método cuasi-harmónico tenía un importante inconveniente asociado a la necesidad de transformar el sistema de coordenadas a coordenadas internas, y a la consecuente aproximación que hay que hacer en el Jacobiano (en el caso de la transformación a coordenadas internas, su determinante es un productorio de distancias de enlace y senos de ángulos de enlace).⁽⁴⁴⁾ La integral anterior es fácil de resolver analíticamente usando (1.1), pero si hacemos un cambio de variable, entonces es muy probable que no seamos capaces resolver el problema analíticamente debido al Jacobiano. La necesidad de ese cambio de variable viene del hecho de que cuando se elimina artificialmente la traslación del centro de masas, lo que es necesario para obtener convergencia de resultados en el tiempo, estamos provocando que la matriz de covarianzas sea singular y, por tanto, su determinante sea nulo. ⁽³⁰⁾ Evidentemente esto es un grave problema si queremos utilizar la ecuación (1.4).

1.1.2 Método de Schlitter

Para superar el inconveniente antes señalado, Schlitter ⁽³⁰⁾ propuso una aproximación empírica a la entropía en la que, a grandes rasgos, suma a la matriz de covarianzas una matriz diagonal haciendo la suma no singular. En concreto, comenzando con la entropía del oscilador armónico cuántico mono-dimensional

$$S_{HO} = k_B \left[\frac{\alpha}{e^\alpha - 1} - \ln(1 - e^{-\alpha}) \right] \quad (1.5)$$

de frecuencia ω y $\alpha = \hbar\omega/k_B T$, el autor propuso usar el teorema de equipartición de la energía con la varianza muestral clásica $\langle x^2 \rangle_c$ y aproximar en el límite clásico ($\alpha \rightarrow 0$) la expresión anterior a:

$$S'_{HO} = \frac{1}{2} k_B \ln \left(1 + \frac{k_B T e^2}{\hbar^2} m \langle x^2 \rangle_c \right). \quad (1.6)$$

Aunque parezca la transformación anterior un salto muy grande, se puede comprobar que no lo es tanto si sustituimos las exponenciales por sus respectivos infinitésimos equivalentes (cosa que por otro lado Schlitter nunca aclaró). De modo que el contenido injustificado de la ecuación (1.6) está básicamente concentrado en el sumando “1” dentro del logaritmo. Tal y como señala el autor, la nueva expresión tiene dos propiedades esenciales: por un lado se anula cuando $\alpha \rightarrow \infty$, lo que estaría de acuerdo con un comportamiento cuántico y, por otro lado, proporciona el valor correcto de entropía en el límite clásico $\alpha \rightarrow 0$.(30) Finalmente, la expresión se puede generalizar para un sistema de múltiples átomos como

$$S'_{HO} = \frac{1}{2} k_B \sum_i \ln \left(1 + \frac{k_B T e^2}{\hbar^2} \langle q_{ii}^2 \rangle_c \right) = \frac{1}{2} k_B \ln \left(\prod_i \left[1 + \frac{k_B T e^2}{\hbar^2} \langle q_{ii}^2 \rangle_c \right] \right),$$

donde $\langle q_{ii}^2 \rangle_c$ es la varianza muestral clásica de los auto-vectores de la matriz $\boldsymbol{\sigma}' = \mathbf{M}\boldsymbol{\sigma}$ siendo \mathbf{M} la matriz de masa.

El elemento más relevante en esta aproximación es que, hasta entonces, la única forma que existía para evaluar la entropía *absoluta* era mediante los modos normales de vibración. Gracias al método de Schlitter, por primera vez fue posible realizar estimaciones de entropías absolutas (en lugar de relativas) de una forma diseñada especialmente para sistemas con un gran número de mínimos locales.

1.1.3 Análisis Cuasi-harmónico: Segunda Aproximación

Poco años después, sin embargo, el análisis cuasi-harmónico fue revisado nuevamente por Karplus y colaboradores.(12) La nueva versión del QHA (*Quasi Harmonic Analysis*) representaba un cambio cualitativo con respecto a la idea original. Básicamente se trata de construir una matriz Hessiana directamente a partir de la matriz de covarianzas, es decir $(\mathbf{H})_{ij} = k_B T (\boldsymbol{\sigma}^{-1})_{ij}$, siguiéndose a continuación el mismo procedimiento que en el análisis de modos normales. Tal y como los autores señalan, el nuevo método da el valor exacto (bajo la consideración de que la hipótesis cuasi-harmónica es válida) de la magnitud que el método de Schlitter trata de aproximar. (12)

A pesar de las mejoras, el método QHA presenta aún algunos inconvenientes. Por un lado, debido a que la cuantificación de la correlación se hace a través de las covarianzas, únicamente las correlaciones lineales son tenidas en cuenta. Además, la superficie de energía potencial del sistema es formalmente suavizada a tal extremo que, aun teniendo en la práctica un número elevado de mínimos, se aproxima por otra superficie con uno sólo. De este modo, se ignora cualquier tipo de anarmonicidad incluyendo la multimodalidad (múltiples máximos en la FDP).

Finalmente, no hay una forma correcta (a no ser usando coordenadas internas) de separar la rotación molecular de los movimientos internos, habiendo incertidumbres de hasta 80 J/(K mol) asociadas a la arbitrariedad en la selección de los átomos que permiten realizar un ajuste mínimo-cuadrático de las distintas configuraciones sobre una estructura de referencia. (45) Además, debido a esta incertidumbre, la entropía incluirá al menos errores aleatorios, pero también un inherente error sistemático. La selección del sistema de coordenadas sin duda influye sobre el resultado final (46) y la mejor estimación la dará (para cada sistema en concreto) el sistema de coordenadas que mejor se adapte a él, es decir, que minimice su entropía. La tarea de buscar el mejor sistema de coordenadas para cada sistema es inabordable y por tanto no podemos garantizar la cancelación de errores sistemáticos de este tipo, como tampoco podemos garantizar que se cancelen los errores debido a la selección de una forma funcional para la FDP incorrecta. En general, QHA siempre nos dará una cota superior de la entropía exacta, (46) debido en parte a que dado un sistema con una matriz de covarianzas $\boldsymbol{\sigma}$, la distribución que maximiza la entropía es precisamente la distribución normal.(47)

La aproximación QHA sigue siendo en la actualidad la más conocida y la más utilizada para la estimación de entropías en biomoléculas. De hecho, constituye el punto de partida de dos métodos muy similares propuestos por Baron (13, 38) y por Numata, (37) donde los autores intentan corregir *a posteriori* los efectos de la anarmonicidad en los modos cuasi-harmónicos y de las correlaciones supralineales entre ellos. En ambos casos, y a pesar de que la entropía QHA es cuántica, la corrección que se hace es clásica (en el caso de Barón) y configuracional (o estadística) en el caso de Numata. La estrategia para corregir la anarmonicidad pasa por calcular las diferencias entre la contribución a la entropía (clásica o estadística) que realmente hace el modo anarmónico y la contribución teórica considerando que la hipótesis harmónica es cierta. Por otro lado, ambos métodos proponen corregir el efecto de las correlaciones supralineales mediante correcciones de información mutua entre los modos cuasi-harmónicos linealmente no correlacionados. Según las observaciones que ha hecho Barón en péptidos modelo, el efecto de la anarmonicidad es muy pequeño mientras que el de las correlaciones supra-lineales es bastante alto, de hecho éste último representa más del 99% de la corrección que se realiza a la entropía.(38)

Un último método, que incluimos en este apartado por utilizar la matriz de covarianzas, fue propuesto recientemente por Brüschweiler y colaboradores.(33) En este método se separan los denominados grados de libertad *duros*, asociados a vibraciones de valencia y ángulos de enlace, de los *blandos* que están relacionados con los ángulos diedros. Con estos últimos se construye la matriz de covarianzas, y la entropía *configuracional* asociada a ellos explicaría los cambios de entropía del sistema. Esta vez, en la diagonalización de la matriz de covarianzas, no interesa para nada los valores propios; la entropía configuracional asociada a los diedros se calcula directamente sobre las distribuciones de los modos (auto-vectores de la matriz) libres de correlaciones lineales entre sí. Según los autores, y esto nos ha llamado la atención, en este método no es recomendable hacer correcciones *a posteriori* de información mutua como las propuestas por Numata y Baron, (13, 37-38) ya que se obtienen peores resultados debido a un aumento en las incertidumbres estadísticas. El método fue aplicado con éxito a un di-péptido de Alanina, sin embargo, se constataron grandes dificultades en cuanto a la convergencia de los valores de entropía durante el estudio de un péptido con un tamaño significativamente mayor (35 residuos, HP35).(33)

1.1.4 Método NN (*Nearest-Neighbor*)

En vez de asumir una forma funcional para la densidad de probabilidad $P(\mathbf{q})$, tal y como hacen los métodos QHA, también es posible estimar la densidad de probabilidad sin ninguna consideración impuesta sobre su forma funcional. Esta clase de estimaciones (no paramétricas) son por tanto mucho más robustas por definición. Sin embargo, la falta de información acerca de la forma funcional de la función de densidad de referencia, debe ser compensada mediante la realización de mucho más muestreo. El estimador *NN* (*Nearest Neighbor*) (40-41) es un ejemplo clásico en el que tanto la densidad como la entropía son estimados no paramétricamente de una forma simple.

Siendo $\mathcal{Q} = \{q_1, q_2, \dots, q_n\}$ el conjunto de n observaciones (por ejemplo de dinámica molecular) de nuestro sistema d -dimensional, la idea es estimar en cada punto $q_i \in \mathbb{R}^d$ la densidad como

$$\hat{\rho}(q_i) = \frac{1}{n} \frac{k}{V_d(R_{i,k})}, \quad (1.7)$$

donde $V_d(R_{i,k}) = \pi^{d/2} R_{i,k}^d / \Gamma(d/2 + 1)$ es el volumen de una esfera d -dimensional de radio $R_{i,k}$ centrada en q_i (el radio de dicha esfera se escoge tal que los k primeros vecinos estén contenidos en la misma). El valor de k debe ser lo suficientemente grande como para suavizar la función estimada de la FDP y lo suficientemente pequeño para que la estimación describa bien localmente a la FDP (valores típicos van desde uno hasta cinco). Nótese que la expresión (1.7) es bastante intuitiva, ya que k/V_d es la densidad de puntos y el pre-factor $(1/n)$ se introduce para normalizar $\hat{\rho}(q_i)$.

Recordando que $S = k_B E(-\ln P(\mathbf{q}))$, nuestra estimación puede ser formulada sustituyendo el valor esperado muestral $E(\cdot)$ por una simple media aritmética:

$$\hat{S}_k = -\frac{1}{n} \sum_{i=1}^n k_B \ln \left(\frac{1}{n} \frac{k}{V_d(R_{i,k})} \right).$$

Sin embargo, se ha comprobado que el valor esperado para un muestreo infinito de este estimador no es el valor real de la entropía configuracional S_{Config} , sino que es igual a $(S_{Config} + L_{k-1} - \ln k - \gamma)$, donde $L_0 = 0$, $L_{j \neq 0} = \sum_{i=1}^j 1/i$ y $\gamma = 0.5772\dots$ es la constante de Euler.(40) En consecuencia, el correspondiente estimador asintóticamente no sesgado sería:²

$$\hat{S}_k = -\frac{1}{n} \sum_{i=1}^n k_B \ln \left(\frac{1}{n} \frac{k}{V_d(R_{i,k})} \right) - (L_{k-1} - \ln k - \gamma). \quad (1.8)$$

La gran ventaja del estimador NN es que puede ser fácilmente implementado en combinación con otras metodologías, como por ejemplo el método MIE,(39) métodos de *clustering*,(41) e incluso la propia QHA.(37) Además, da mejores resultados que los métodos de estimación por histograma, especialmente cuando la dimensionalidad del problema aumenta, aunque en cualquier caso está normalmente limitado a menos de 10 dimensiones. Por otro lado, es un método mucho más costoso que el clásico histograma de escalado lineal, ya que determinar el radio de la hiper-esfera que contiene k puntos tiene un escalado cuadrático en n ($O(n^2)$).

Este método, tal y como se ha presentado, es sólo aplicable a sistemas de muy baja dimensionalidad (difícilmente se llega a 12 lo que equivale a ¡4 átomos!). Como primera aproximación para subsanar parcialmente esta limitación, Hnizdo y col.(41) propusieron utilizar un criterio de *clustering*, donde las coordenadas moleculares son clasificadas en grupos excluyentes de dimensiones manejables, de tal forma que la dependencia entre las variables pertenecientes a diferentes grupos sea mínima. La entropía total es entonces aproximada por la suma de las entropías estimadas para cada grupo. Sin embargo, un año después, el mismo autor decidió abandonar esta propuesta y

² Los sesgos de los estimadores aparecen con frecuencia en estadística. Por ejemplo, es fácil verificar que la media muestral es siempre un estimador insesgado de la media poblacional, mientras que la varianza muestral proporciona un valor esperado que es $n^{-1}(n-1)$ veces la varianza poblacional σ^2 (esto justifica la utilización de la cuasi-varianza).

decantarse por combinar el método *NN* con otro método no paramétrico bien conocido (*Mutual Information Expansion*) (39) que describiremos con más detalle a continuación.

1.1.5 Método MIE (*Mutual Information Expansion*)

El método MIE (39, 44, 48) es una expansión sistemática de la entropía en la que se van añadiendo progresivamente correlaciones de órdenes superiores, a través de las funciones de información mutua hasta el orden deseado. De este modo, la dimensión del problema puede reducirse hasta unos límites manejables. De aquí en adelante, si no se aclara lo contrario, cuando hablemos de la función de información mutua MI (*Mutual Information*) nos referiremos a la de orden dos (dos variables). A diferencia de las covarianzas, la función MI caracteriza una dependencia general, es decir, captura toda la correlación de cualquier naturaleza.

La MI entre dos variables q_i y q_j se define en términos de entropía como $I(q_i, q_j) = S(q_i) + S(q_j) - S(q_i, q_j)$, donde $S(q_i)$ y $S(q_j)$ son las entropías marginales de ambas variables y $S(q_i, q_j)$ es la entropía de la variable bidimensional (q_i, q_j) evaluada normalmente con la expresión de Shannon (ecuación (1.2)). Análogamente, se define la información mutua entre k variables como

$$I_k(q_1, \dots, q_k) = \sum_{m=1}^k (-1)^{m+1} \sum_{\substack{\mathcal{J} \subset \{q_1, \dots, q_k\} \\ |\mathcal{J}|=m}} S(\mathcal{J}), \quad (1.9)$$

donde el símbolo $|\cdot|$ se utiliza para la cardinalidad, es decir, el número de elementos del conjunto. Para un m dado, la sumatoria sobre \mathcal{J} corre sobre todos los posibles $\binom{k}{m}$ subconjuntos de $\{q_1, \dots, q_k\}$ que cumplen con la condición $|\mathcal{J}| = m$.

Es común, y de eso trata el método MIE, aproximar la entropía total de un sistema \mathcal{A} de M variables, a una forma truncada hasta un orden dado n , usando su expansión en términos de las funciones MI generalizadas:

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{I} \subset \{q_1, \dots, q_M\} \\ |\mathcal{I}|=k}}^M I_k(\mathcal{I}). \quad (1.10)$$

Esta expansión se acercaría al valor exacto a medida que n se acerca a M que es, por otro lado, la cardinalidad de \mathcal{A} . Sin embargo, en la mayoría de los casos sólo es posible llegar a órdenes menores o iguales que tres.

Killian y col.(44) comprobaron que esta expansión es equivalente a estimar la FDP de todo el sistema. Es por ello que este método (MIE) lo incluimos dentro de los estimadores de densidad no paramétricos. Utilizando coordenadas internas de tipo distancias y ángulos de enlace y ángulos de torsión, estos autores estimaron las entropías configuracionales S_{Config} de sistemas pequeños. También analizaron los cambios de entropía en procesos de unión proteína-ligando, donde se ha comprobado la importancia de una correcta estimación de la entropía y el papel de ésta en los procesos de *binding*.(15, 32)

En cualquier caso, se trata de un método cuya aplicación está limitada a sistemas de pequeño tamaño. Las principales dificultades de este método en términos prácticos son, por un lado, el coste que supone la corrección de la entropía a órdenes superiores. La dificultad no está sólo en el cómputo de estos términos en sí, sino también en los problemas de convergencia asociados, tanto en el orden como en el tiempo. Por otro lado, tal y como hemos comprobado durante el desarrollo de nuestra investigación, las expansiones de tipo MIE no aseguran que elevar el orden implique necesariamente una mejor estimación. Es más, no hay una forma sistemática conocida de saber, dado el muestreo con el que se cuenta, qué truncamiento proporcionaría una estimación óptima.

1.1.6 Método MCSA (*Minimally Coupled Subspace Approach*)

También vale la pena comentar otros métodos relativamente recientes que, aunque no se han extendido, representan otra solución interesante al problema que nos ocupa. Este es el caso del método no paramétrico MCSA,(34, 42) que con alguna excepción, es básicamente una combinación de los métodos que hemos visto hasta el momento. Brevemente, Lange y col. desarrollaron un estimador de densidad tipo *kernel* pero

anisotrópico, basado en el método NN. Con ello en principio pueden integrar dimensiones mayores que el método original NN y éste es el que utilizan en todos los cálculos. En primer lugar, generan lo que llaman subespacios mínimamente acoplados mediante una transformación ortogonal que minimiza la información mutua entre las variables. Las nuevas variables, que no tienen por qué ser ortogonales con el producto escalar habitual, son agrupadas por un método de *clustering* que crea los grupos (*clusters*) mínimamente acoplados. La estimación final de la entropía vendría dada por la suma de las entropías de estos subgrupos. Pero si alguno de estos grupos es de dimensión muy alta ($d > 15$), éste es nuevamente dividido en subgrupos. Estos últimos presumiblemente se correlacionan con todos los de su mismo grupo, y esta correlación se calcula usando las funciones MI.

El principal problema de este método es que implica muchas decisiones empíricas, tales como los criterios de *clustering*, la dimensionalidad máxima de los *clusters* (tamaño muy pequeño que correspondería con 5 átomos) y el orden de truncamiento de las correcciones con las MIFs. Si sumamos todo ello a que no resuelve el problema de la separación entre rotación y movimientos internos, y que sorprendentemente, los autores no muestran resultados de convergencia de método, nos hace observarlo de momento, con cierta cautela.

1.1.7 Métodos HS (*Hypothetical Scanning*)

Para terminar, citaremos otro método no paramétrico diferente en naturaleza a los ya presentados. Se trata del método HS (*Hypothetical Scanning*).⁽⁴⁹⁻⁵⁰⁾ Este método, que se ha utilizado tanto sobre trayectorias MD (HSMD) como MC (HSMC), se basa en reconstruir, paso a paso, la probabilidad de una estructura cualquiera de la trayectoria. Para ello se va reconstruyendo, de cero, el sistema mediante transiciones hipotéticas de probabilidad conocida. Al final, al emplear básicamente probabilidades condicionadas, la probabilidad total de una configuración será igual al producto de las probabilidades de estas etapas. A partir de estas probabilidades y de las energías de cada punto, se estiman tanto la entropía *configuracional* como la propia energía libre.

Hasta ahora, esta metodología ha sido aplicada por sus propios creadores a la estimación de la energía libre en fluidos como el agua y el argón⁽⁵⁰⁾ y en modelos peptídicos,

tanto en vacío como en disolvente explícito. (49) Más recientemente, la han aplicado al estudio de un fragmento modesto de una proteína en disolvente explícito.(16) Sin embargo, aún se está muy lejos de aplicar esta metodología a los grandes sistemas moleculares que son nuestro objetivo en este trabajo.

1.1.8 Nuestra Propuesta para el Cálculo de Entropías Absolutas

La esencia de nuestra propuesta, que se detallará en el Capítulo de Resultados, descansa en una conocida partición de la entropía total.(3, 17, 51) De este modo, la entropía total de un sistema molecular, excluyendo la traslación y la rotación del sistema como un todo, se puede estimar como la suma $S_{tot} = \langle S_{vib} \rangle + S_{conform}$, donde $\langle S_{vib} \rangle = \sum_i p_i S_{vib,i}$ es la entropía vibracional promediada sobre los diferentes mínimos de la superficie de energía potencial y $S_{conform}$ es la entropía asociada a la incertidumbre de pertenecer a uno u otro mínimo (confórmero). La gran ventaja de esta aproximación es que $\langle S_{vib} \rangle$ se puede evaluar por métodos convencionales, como el clásico análisis de modos normales. Mientras tanto, la entropía restante $S_{conform}$ está asociada a una variable aleatoria discreta “el estado conformacional”, que la hace *en principio* relativamente sencilla de calcular (no necesitamos resolver integrales, simplemente sumatorias). Mediante un procedimiento que se detallará en el Capítulo 2, nuestro método discretiza los ángulos diedros en estados conformacionales para el posterior cálculo. Finalmente, aprovechando la redundancia de las expansiones MIE, transformamos la dependencia de éstas en el orden en una dependencia con el *cut-off*. Esto supone un salto cuantitativo en la eficiencia del método y facilita la implementación de correcciones que nos ayuden a eliminar la falsa correlación.

1.2 Cálculos de Energía en Biomoléculas basados en Fragmentos

Dado el coste de los cálculos mecánico-cuánticos (QM, *Quantum Mechanical*) en sistemas de gran tamaño como las biomoléculas y la necesidad de utilizar niveles de cálculo precisos para obtener buenas estimaciones de la estabilidad y reactividad de estos sistemas, la idea de representar la energía total del sistema como combinación de la energía de sus fragmentos es una alternativa que se ha venido considerando desde hace bastante tiempo. En este apartado haremos una revisión de las principales metodologías desarrolladas con este objetivo en las últimas décadas. Discutiremos primero los métodos basados en expansiones de varios cuerpos, incluyendo algunas variantes que contienen implícitamente los efectos de orden superior. Comentaremos también el método llamado KEM (*Kernel Energy Method*), que resulta ser esencialmente un caso particular de MBE (*Multi-Body Expansion*). Finalmente revisaremos otros métodos que aproximan la energía QM de un sistema dado como combinación de las energías de los fragmentos apoyándose en una justificación de naturaleza termodinámica.

Veremos que los protocolos basados en justificaciones termodinámicas, se pueden considerar como formas truncadas de una expansión más general de tipo MBE. Sin embargo, estos protocolos de tipo termodinámico son conceptualmente más simples y pueden ser fácilmente implementados con un bajo coste computacional. Precisamente, en el Capítulo de Resultados presentaremos nuestro propio método basado en consideraciones termodinámicas, y que creemos que puede resultar especialmente útil para el estudio de biomoléculas. Se mostrará la precisión del método en cálculos de prueba y, una vez validado, será aplicado a un modelo de colágeno para estimar su energía libre de formación a partir del estado monomérico.

1.2.1 Método MBE (*Multi-Body Expansion*)

El método MBE (52-53) es el análogo en energía, al denominado método MIE para el cálculo de entropía. Ambos guardan similitud con el conocido principio de inclusión-exclusión de la teoría de conjuntos y de la teoría de probabilidad.(54-55) El método MBE evalúa la energía total de un sistema como una suma de términos que va

incluyendo progresivamente los efectos de dos-, tres-, ..., n -cuerpos. Este método puede ser generalizado para cualquier tipo de sistemas, evaluándose la energía con series de contribuciones, independientes del resto de la estructura y perfectamente transferibles. (53)

En este formalismo general, la energía total de un sistema de M partículas (átomos, moléculas, o fragmentos unidos covalentemente) podríamos expresarla como

$$E_M(A_1, A_2, \dots, A_M) = \sum_{N=1}^M E^{(N)}(A_1, A_2, \dots, A_M), \quad (1.11)$$

donde $A_i = \{\mathbf{q}_i, \eta_i\}$ contiene la información acerca de las coordenadas (\mathbf{q}_i) y de la naturaleza (η_i) de la partícula i -ésima. Como la numeración de las partículas es arbitraria, la forma funcional debe ser tal que E_M sea invariante frente a cualquier permutación del tipo $A_i \leftrightarrow A_j$. Por otro lado, cada término $E^{(N)}$ se puede calcular como una suma múltiple de potenciales de N -cuerpos, es decir:

$$E^{(N)} = \sum_{m_1 < \dots < m_N}^M V^{(N)}(A_{m_1}, A_{m_2}, \dots, A_{m_N}). \quad (1.12)$$

Nótese que la ecuación (1.12), y por tanto la (1.11), expresan la energía total en términos de los potenciales de N -cuerpos que no hemos definido aún. En la práctica, sin embargo, necesitamos obtener estos potenciales $V^{(N)}$ a partir de cálculos hechos sobre diferentes subsistemas. La relación general entre $V^{(N)}$ y las energías de los distintos subsistemas se puede obtener mediante la inversión de Möbius definida en la teoría de números. (53) La expresión general que se obtiene es:

$$V^{(N)}(A_1, A_2, \dots, A_N) = \sum_{L=1}^N (-1)^{N-L} \sum_{m_1 < \dots < m_L}^N E(A_{m_1}, A_{m_2}, \dots, A_{m_L}). \quad (1.13)$$

En la ecuación anterior, $E(A_{m_1}, A_{m_2}, \dots, A_{m_L})$ evalúa la energía del conjunto formado por L fragmentos $\{m_1, \dots, m_L\}$. De hecho, la ecuación (1.13) constituye la única definición de

los potenciales de N -cuerpos, que los hace independientes de la estructura y perfectamente transferibles, porque en la evaluación de cada uno de ellos no se incluye ninguna información acerca de su entorno.

El verdadero significado de la expansión (1.13) se puede visualizar mejor si mostramos de forma explícita los primeros términos $E^{(N)}$. Para el caso de la primera aproximación $E^{(1)}$, se recoge simplemente la suma de las energías de las partes en las que hemos dividido el sistema.

$$E^{(1)} = \sum_{m_1=1}^M V^{(1)}(A_{m_1}) = \sum_{m_1=1}^M E(A_{m_1}) \quad (1.14)$$

El segundo término $E^{(2)}$ corregiría, en una segunda aproximación, el error que cometemos al considerar que las energías son aditivas. Se trata simplemente de la suma de las interacciones de pares. Por tanto, la suma de los dos primeros términos $E_M \approx E^{(1)} + E^{(2)}$ sería lo que conocemos como *pairwise approximation*

$$E^{(2)} = \sum_{m_1 < m_2}^M V^{(2)}(A_{m_1}, A_{m_2}) = \sum_{m_1 < m_2}^M \left[E(A_{m_1}, A_{m_2}) - E(A_{m_1}) - E(A_{m_2}) \right]. \quad (1.15)$$

Análogamente, y para terminar, la contribución de $E^{(3)}$ sería la energía adicional debido al efecto de tres cuerpos que no queda capturada a partir de una aproximación a pares, es decir:

$$V^{(3)} = \sum_{m_1 < m_2 < m_3}^M \left[E(A_{m_1}, A_{m_2}, A_{m_3}) - E(A_{m_1}) - E(A_{m_2}) - E(A_{m_3}) - V^{(2)}(A_{m_1}, A_{m_2}) - V^{(2)}(A_{m_1}, A_{m_3}) - V^{(2)}(A_{m_2}, A_{m_3}) \right]. \quad (1.16)$$

1.2.2 Método KEM (*Kernel Energy Method*)

En este punto es conveniente simplificar la notación usada en las ecuaciones MBE reemplazando $E(A_{m_1}, A_{m_2}, \dots, A_{m_L})$ (la energía de un subsistema de L fragmentos) por una nomenclatura de tipo $E_{ijk\dots}$. De esta forma, la aproximación a pares quedaría:

$$E_M = \sum_{i=1}^M E_i + \sum_{i=1}^M \sum_{j=i+1}^M (E_{ij} - E_i - E_j). \quad (1.17)$$

Recientemente, el método denominado KEM ha sido utilizado para calcular energías QM de grandes biomoléculas (56-60) haciendo una partición exclusiva en fragmentos del sistema llamados *kernels*. La mayoría de las aplicaciones reportadas de este método calculan la energía total mediante la expresión

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M-2) \sum_{i=1}^M E_i. \quad (1.18)$$

Se puede comprobar fácilmente, y así lo hemos hecho,(61) que la fórmula anterior no es más que el método MBE hasta segundo orden, es decir, la ecuación (1.17). Sin embargo, hay que señalar que las aplicaciones del método KEM reportadas en sistemas covalentes de gran tamaño saturan los enlaces que conectan a los *kernels* entre sí con átomos de hidrógeno, como paso previo al cálculo de la energía. La presencia de estos hidrógenos induce errores, ya que su introducción es ajena al método MBE. Aunque, por otro lado, si los fragmentos son lo suficiente grandes y se reduce su número, el error asociado puede ser razonablemente pequeño. Lo mismo ocurre si se van incluyendo progresivamente términos de órdenes superiores. En este caso, el error se va reduciendo al ir considerándose grupos de fragmentos cada vez mayores, que van describiendo mejor el entorno de cada fragmento.(60)

1.2.3 Métodos Corregidos con un Entorno Electroestático

En principio, la aproximación de pares definida en la ecuación (1.17) puede no ser suficiente para describir adecuadamente la energía total de conjuntos moleculares en fase condensada. Lamentablemente, el cálculo de términos de orden superior que

pueden jugar un papel importante es extremadamente costoso. Para superar esta limitación de las expansiones de segundo orden con un coste computacional razonable, algunos autores han propuesto calcular las energías de los fragmentos individuales (E_i) y de sus pares (E_{ij}) teniendo en cuenta el campo electrostático generado por el resto del sistema. (62-67) Por ejemplo, en el método FMO (*Fragment Molecular Orbital*), las energías de los diferentes fragmentos son calculadas de manera iterativa resolviendo los Hamiltonianos de los fragmentos e incluyendo de modo efectivo el efecto de los electrones de los $M-1$ fragmentos restantes y de sus núcleos.(63) Las energías FMO resultantes son combinadas usando las ecuaciones MBE de orden 2-3 para derivar la energía total. Una alternativa similar para fragmentos no enlazados covalentemente es el método *Electrostatic-Embedded* MBE (EE-MBE).(65-67) La energía, en este caso, se calcula en presencia del campo electrostático generado por las cargas puntuales del resto de fragmentos, observándose una mejora significativa de las estimaciones a orden dos y tres en modelos de *clusters* de agua, en comparación con la formulación estándar MBE. (65)

1.2.4 Método MFCC (*Molecular Fractionation with Conjugate Caps*) y Método MTA (*Molecular Tailoring Approach*)

El método llamado MFCC estima la energía total a partir de fragmentos sobre una base intuitiva y/o termoquímica. Fue originalmente propuesto para calcular energías de interacción QM entre una proteína y un ligando pequeño,(68) pero el método ha sido ya expandido para predecir la energía QM total de toda una proteína.(69) En esta aproximación la proteína es dividida en fragmentos $A_i = (-C_\alpha HR_i - CO - N_{i+1}H -)$, donde R_i es la cadena lateral del residuo i -ésimo y N_{i+1} es el átomo de nitrógeno del *backbone* del residuo siguiente. En lugar de saturar las valencias con Hidrógenos, como sucede en el método KEM, se utilizan dos “*conjugate caps*”, “ NH_2- ” y “ $-C_\alpha H_2 R_{i+1}$ ” que se colocan respectivamente en los átomos $C_{\alpha,i}/N_{i+1}$ de cada fragmento A_i . La energía total de la proteína de M residuos es primeramente aproximada por una suma de las energías de los fragmentos con sus valencias saturadas y, seguidamente, se sustrae las energías de los $NH_2 - C_\alpha H_2 R_{i+1}$ “*conjugate caps*”. Esta aproximación, que podemos llamar de primer orden, es posteriormente corregida con un término ($\delta E^{(2)}$) que da

cuenta de las interacciones de pares entre fragmentos no vecinos. La expresión final quedaría:

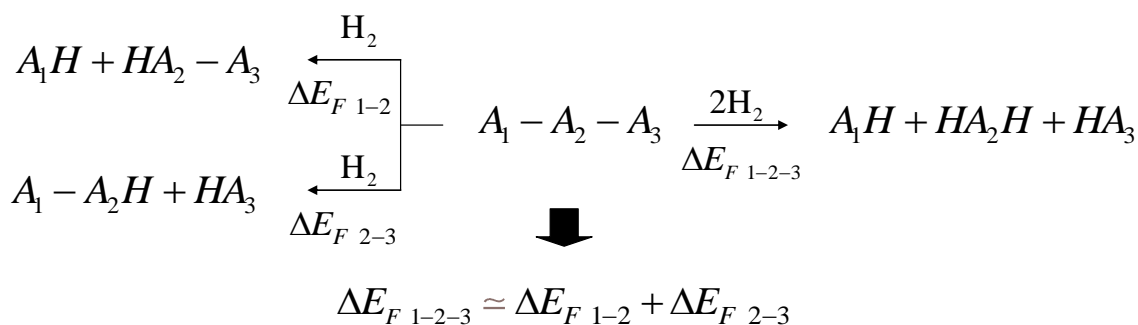
$$E_M = \left[E(A_1 - C_\alpha R_2 H_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - R_{i+1} C_\alpha H_2) + E(NH_2 - A_M) \right] - \left[\sum_{i=1}^{M-1} E(NH_2 - R_{i+1} C_\alpha H_2) \right] + \delta E^{(2)} \quad (1.19)$$

Para calcular $\delta E^{(2)}$, las valencias de los fragmentos son saturadas con hidrógeno al igual que en método KEM.(56) Alternativamente se ha propuesto otra variante de este método, donde al igual que en EE-MBE, las energías QM de los fragmentos son calculadas en un entorno de cargas puntuales representando al resto de residuos. (70)

Por otro lado, el método MTA (*Molecular Tailoring Approach*) (71) divide al sistema en fragmentos que se superponen, estimando la energía total como la suma de estos fragmentos menos la suma de las energías de las intersecciones. Nótese que el método MFCC se puede considerar como un caso particular de MTA, dado que los fragmentos MFCC saturados son equivalentes a los fragmentos MTA que se superponen, y las “tapas conjugadas” serían las intersecciones MTA.

1.2.5 Método de Fragmentación Sistemática

Como veremos, el método MFCC puede ser justificado por medio de simples argumentos termoquímicos sobre la base de una fragmentación formal del sistema. De hecho, aproximaciones termoquímicas para calcular la energía de grandes moléculas basadas en fragmentos han sido utilizadas sistemáticamente por Collins y col. (72) El razonamiento básico detrás de la generalización propuesta por Collins está resumido en el Esquema 1, que muestra un sistema molecular genérico compuesto por tres fragmentos (A_1 - A_2 - A_3) que puede ser formalmente “cortado” de tres formas diferentes.


Esquema 1

La aproximación clave hecha por Collins es que la energía de reacción en el proceso de fragmentación total de $A_1-A_2-A_3$ ($\Delta E_{F\ 1-2-3}$), se aproxima a la suma de energías de reacción correspondientes a las dos fragmentaciones por separado $\Delta E_{F\ 1-2}$ y $\Delta E_{F\ 2-3}$. La consecuencia inmediata es la siguiente:

$$E_{123} = E_{12} + E_{23} - E_2. \quad (1.20)$$

Collins y col. han empleado tanto la topología química como consideraciones sobre costes computacionales, con el fin de seleccionar el mejor sitio de fragmentación de una molécula determinada. El objetivo es que el fragmento A_2 resultante sea: (a) lo suficientemente grande como para poder despreciar la interacción entre A_1 y A_3 , y (b) lo suficientemente pequeño como para calcular la energía del fragmento A_1-A_2H utilizando métodos QM. Si el fragmento que acompaña a $HA_2 - A_3$ es demasiado grande, el protocolo de fragmentación que se define en el Esquema 2 se aplica iterativamente, hasta que todos los fragmentos producidos se puedan describir mediante mecánica cuántica. En última instancia, en este enfoque termoquímico la energía total se aproxima por una combinación lineal de las energías de los fragmentos, cuya forma exacta depende de la topología del sistema y de consideraciones computacionales.

1.2.6 Observaciones Generales sobre los distintos Métodos

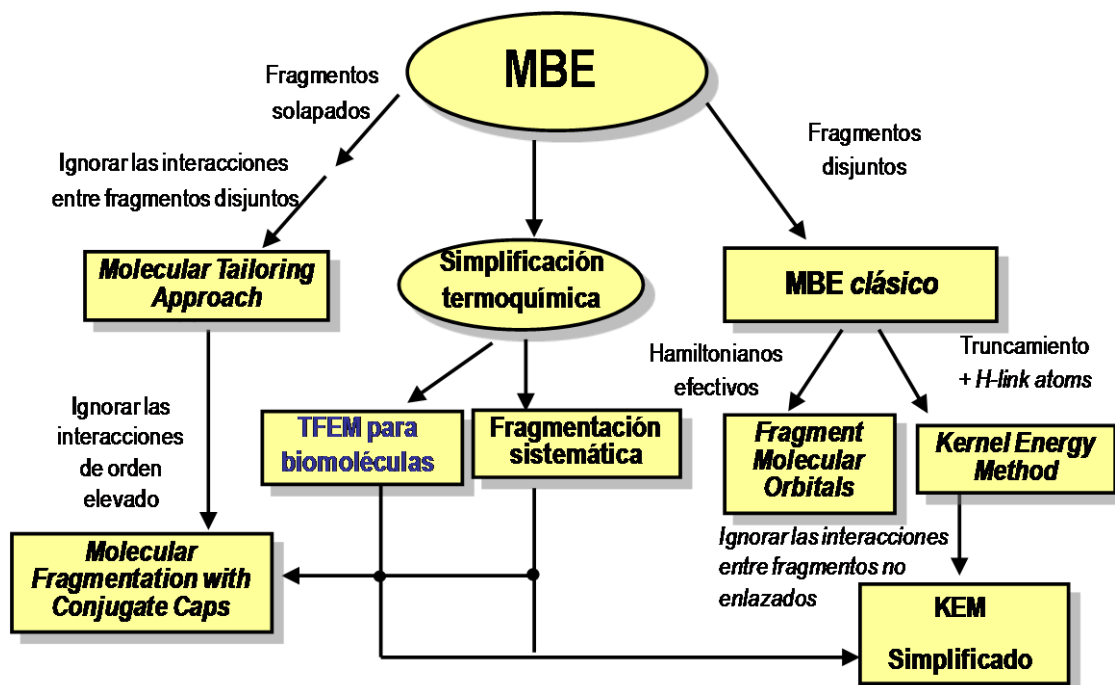
Hemos mostrado un conjunto de aproximaciones con las que tenemos una panorámica de la variedad de métodos que se han publicado sobre este tema, comentando aquellos que hemos considerado más relevantes. Evidentemente, existen más métodos como el reciente *Molecules in Molecules* (MIM)(73) que es una extensión natural del popular

método ONION propuesto originalmente por Morokuma.(74) En cualquier caso, este método MIM es quizás el único que no puede considerarse como un caso particular de MBE, que como veremos a continuación es la piedra angular de las aproximaciones que se han descrito.

Puede ser interesante recordar que la ecuación MBE para orden N conduce a una combinación lineal de las energías de los fragmentos. De hecho, el método de Collins y col. de fragmentación sistemática se puede generar directamente del método MBE, despreciando todos los términos de interacción más allá de segundo orden y usando un criterio de topología adicional para despreciar un gran número de contribuciones de segundo orden. Por otra parte, la inclusión de los átomos de Hidrógeno saturando las valencias expuestas de los fragmentos extraídos de un sistema covalente, hace el método Collins de fragmentación casi idéntico a la versión simplificada del método KEM, en el que sólo se consideran los dobles *kernels* unidos químicamente.(57) Así, una vez que el esquema de fragmentación se ha aplicado, en ambos métodos se calculan los mismos términos de energía.

Resumiendo, el formalismo MBE establece el marco general para el desarrollo computacional de estrategias dirigidas a la evaluación de la energía total de los grandes sistemas a partir de energías de subsistemas (ver Esquema 2). De este modo, la revisión bibliográfica realizada durante el desarrollo de nuestro trabajo puso de manifiesto un hecho que, en gran medida, fue ignorado en algunos de los trabajos anteriores. Así, el método FMO, las diversas fórmulas KEM, y la expresión MFCC con las interacciones de pares se pueden clasificar como técnicas MBE que incluyen los efectos de N -cuerpos. Del mismo modo, la fragmentación sistemática del método de Collins puede ser generada directamente desde el MBE, al pasar por alto todos los potenciales de interacción MBE más allá de segundo orden y con el uso criterios topológicos para despreciar un gran número de contribuciones de segundo orden. También podemos ver en el Esquema 2 que, la saturación de valencias con átomos de Hidrógeno hace al método de fragmentación sistemática de Collins idéntico a la versión simplificada del método de KEM, en el que sólo los *kernels* unidos químicamente son considerados. Por otra parte, el método MFCC se puede considerar como un caso particular del formalismo MTA, ya que los fragmentos MFCC saturados son equivalentes a la superposición de fragmentos de MTA y los *conjugate caps* de MFCC corresponderían a

las intersecciones de fragmentos en el enfoque MTA. Sin embargo, mientras que los fragmentos MFCC son construidos para hacer superposiciones simples (es decir, cada átomo sólo puede ser parte de uno o dos fragmentos), el método MTA admite superposiciones más complejas de N fragmentos.



Esquema 2

1.2.7 Nuestra Propuesta para el Cálculo de Energías basadas en Fragmentos

En esta Tesis, hemos diseñado y aplicado un método que calcula con razonable precisión la energía QM de un sistema de gran tamaño a partir de las energías de sus fragmentos.^(61, 75) Nuestra aproximación se basa en considerar la energía electrónica como una función de estado, de modo que podemos calcular la energía involucrada en la ruptura “formal” de una molécula en fragmentos que a su vez puedan ser tratados íntegramente con métodos QM y derivar, entonces, con la misma calidad la energía de “toda” la molécula haciendo uso de un ciclo termodinámico. La “fragmentación” puede ser cualquier reacción que “corte” la molécula y, al tratarse de una reacción formal, puede combinarse fácilmente con cualquier modelo de disolvente. Finalmente obtendremos la energía total estimada como suma y resta de energías de fragmentos.

1.3 La Triple Hélice de Colágeno

La temática central de esta tesis es el desarrollo de propuestas para una mejor estimación de la entropía y la energía en grandes sistemas biomoleculares, cuyo contexto metodológico ha sido presentado en los dos epígrafes anteriores, y su aplicación durante el estudio de la triple hélice de colágeno. El interés del colágeno en el contexto de esta Tesis radica en el importante papel que juega la entropía durante el proceso de formación de la molécula o tropocolágeno, ya que se pasa de una estructura disociada muy flexible a un estado de triple hélice con grandes restricciones conformacionales.(26) Incluso la propia cinética de formación del colágeno está controlada en gran medida entrópicamente.(25) Además, la estructura lineal del colágeno parece también muy adecuada para tratar de estimar la energía QM de toda la molécula a partir de la energía de sus fragmentos, refinando así las energías MM proporcionadas por los campos de fuerza tradicionales. En los siguientes epígrafes resumimos los principales resultados de interés en relación al colágeno.

1.3.1 Relevancia y función del Colágeno

El colágeno es la proteína más abundante en la matriz extracelular de los animales (constituye casi la cuarta parte de su contenido total en proteínas), donde las fibras de colágeno proporcionan al medio extracelular las propiedades mecánicas adecuadas para cada tipo de tejido. De este modo, las fibras de colágeno son un componente fundamental en todos los tejidos de soporte, tales como los huesos, los tendones, los ligamentos, la piel, o los vasos sanguíneos. Pero, además de dar estructura y firmeza, el colágeno interpreta un importante papel funcional ya que proporciona puntos de unión específicos para el resto de moléculas de la matriz. Asimismo, gracias a su capacidad para interactuar con las integrinas (proteínas de membrana que conectan el citoesqueleto de las células con el entorno pericelular), modula todos aquellos procesos relacionados con la adhesión celular (migración, proliferación, diferenciación, etc.).(76-77)

La familia de las proteínas colagénicas es muy compleja y variada, no sólo en cuanto a su secuencia, sino también en cuanto a su organización supramolecular, su distribución en los diferentes tejidos y su función.(78) Existen al menos 27 tipos de colágeno distintos, recogiendo en la Tabla 1.1 algunas de las características más relevantes de los cuatro primeros tipos. Se citan también algunas de las patologías asociadas a la presencia de mutaciones puntuales en la secuencia del colágeno, que generalmente afectan a los tejidos de soporte (huesos, cartílago y vasos sanguíneos), lo que recalca su importante papel estructural.

Tipo	Moléculas	Organización	Tejidos	Ejemplo de patología asociada
I	$[\alpha 1(I)]_2\alpha 2(I); [\alpha 1(I)]_3$	Fibrilar	Piel, huesos, tendones, ligamentos, cornea	Osteogénesis imperfecta I-IV, síndrome de Ehlers-Danlos
II	$[\alpha 1(II)]_3$	Fibrilar	Cartílago	Acondrogénesis II,
III	$[\alpha 1(III)]_3$	Fibrilar	Piel, vaso, intestino, útero	síndrome de Ehlers-Danlos tipo vascular
IV	$[\alpha 1(IV)]_2\alpha 2(IV)$ $\alpha 3(IV)\alpha 4(IV)\alpha 5(IV)$ $[\alpha 5(IV)]_2\alpha 6(IV)$	Redes	Membrana basal	Síndrome de Alport

Tabla 1.1: Algunas características de los primeros cuatro tipos de colágeno

1.3.2 Estructura de la Molécula de Colágeno

Desde un punto de vista estructural, se define una proteína como colagénica si contiene la triple hélice de colágeno de forma mayoritaria en su estructura molecular. Dicho así es de esperar encontrar dominios “no colagénicos” o globulares en los distintos tipos de colágeno, principalmente en los extremos *N*- y *C*-terminal de la proteína. La estructura característica de triple hélice(79-80) está formada por tres cadenas polipeptídicas (Fig.1.1), cada una de ellas con estructura de hélice α levógira y cuya secuencia está desplazada en un residuo entre hebras consecutivas, dando lugar a una triple hélice dextrógira. Cada una de estas cadenas se caracteriza por presentar una secuencia regular del tipo $(X-Y-Gly)_n$, siendo *n* el número de tripletes que típicamente es elevado (mayor de 350 en el caso del colágeno I). Siempre se encuentra un aminoácido glicina en la tercera posición de cada triplete, ya que el empaquetamiento compacto de la triple hélice requiere un residuo estéricamente pequeño cada tres posiciones. Si sustituimos una de estas glicinas por un residuo Alanina, por ejemplo, se observaría la pérdida

parcial de la estructura colagénica con un desenrollamiento local de la triple hélice.(80) Además, las posiciones **X** e **Y** son habitualmente ocupadas por aminoácidos conformacionalmente restringidos, como la Prolina (Pro) o la 4(R)-Hidroxirolina (Hyp). De este modo resulta la denominada secuencia prototípica (Pro-Hyp-Gly)_n, cuya triple hélice es especialmente estable y compacta, presentando un ángulo de giro próximo a los 51.4° por triplete y un paso de hélice (distancia entre carbonos α de posiciones contiguas equivalentes) de aproximadamente 8.6 Å. Por ello se acerca a un modelo de hélice 7₅.

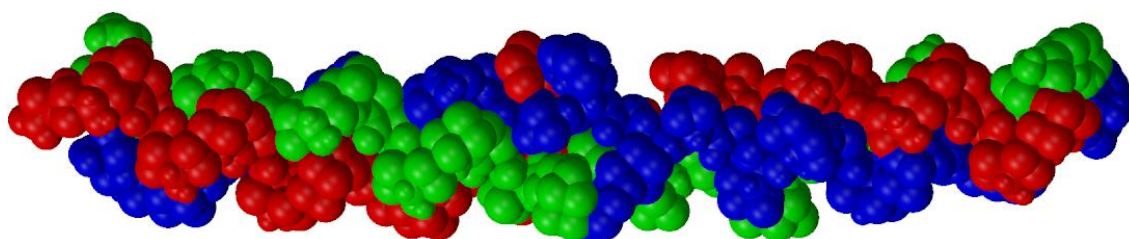


Figura1.1: Fragmento de una triple hélice de colágeno. Las tres cadenas se han coloreado de manera diferente para diferenciarlas. (código PDB: 1BKV).

La estructura en triple hélice del colágeno presenta unos patrones característicos de puentes de hidrógeno(79) entre los grupos amino de los residuos glicina de una cadena polipeptídica y los grupos carbonilo en posición **X** de las cadenas adyacentes (Figura 1.2A). Además, cuando la posición **X** no está ocupada por un imido-ácido, su grupo amino puede conectarse a través de un puente de hidrógeno mediado por una molécula de agua, con el grupo carbonilo de una glicina situada en una cadena adyacente (Figura 1.2B) (81).

Las moléculas de agua en general desempeñan un papel importante en la estructura y la estabilidad de la molécula de colágeno. Los experimentos de difracción de rayos-X en modelos colagénicos revelan que cada triple hélice está rodeada por un cilindro de aguas de hidratación.(80) Hay al menos tres tipos de moléculas de agua(82) químicamente distinguibles ya que presentan diferentes tiempos de relajación (T_2) y, por tanto, diferentes coeficientes de difusión (D). Además, los experimentos de resonancia magnética nuclear (RMN) demuestran que la primera capa de hidratación es

cinéticamente lábil, con tiempos de residencia de que oscilan entre fracciones de nano-segundos hasta varios nano-segundos.(83)

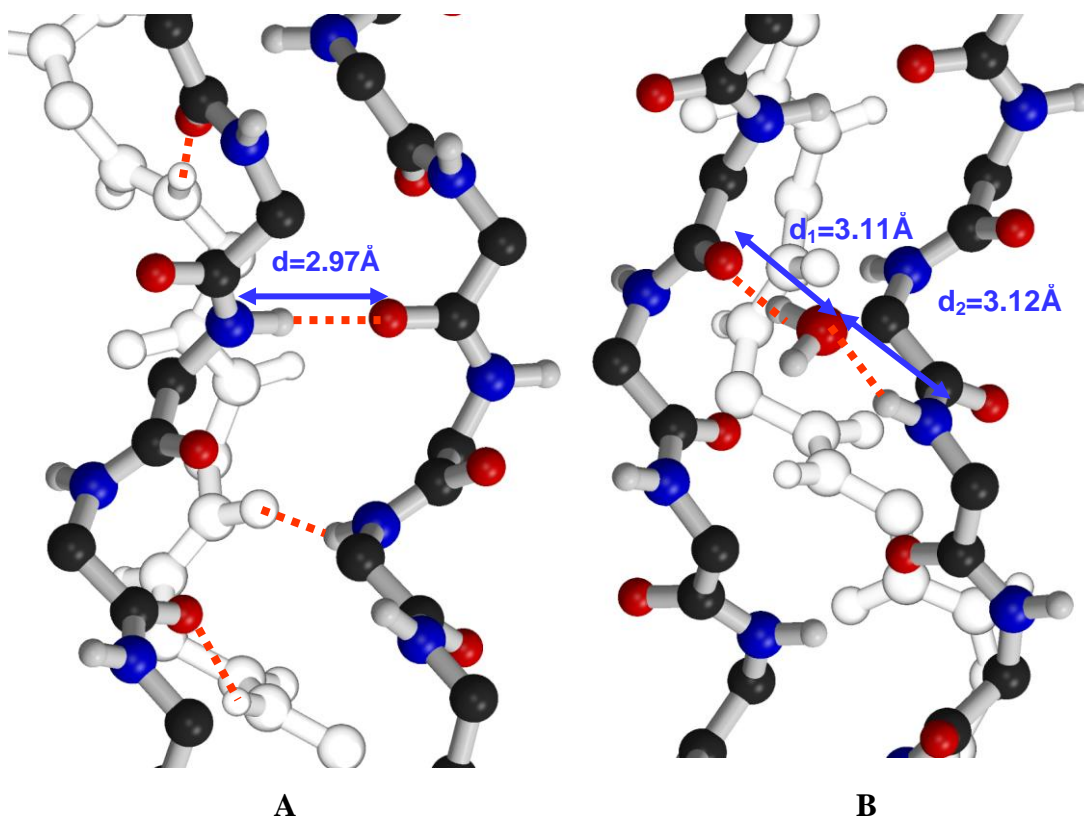


Figura 1.2: Representación de un fragmento del “backbone” o cadena principal de una triple hélice de colágeno. **A:** puentes de hidrógeno directos entre cadenas, donde el N-H de las glicinas interactúa con los grupos C=O en posición X de una cadena adyacente. **B:** puente de hidrógeno mediado por agua, donde los grupos C=O de las glicinas se conectan con grupos N-H en posición X de una cadena adyacente mediante una molécula de agua. (código PDB: 1BKV).

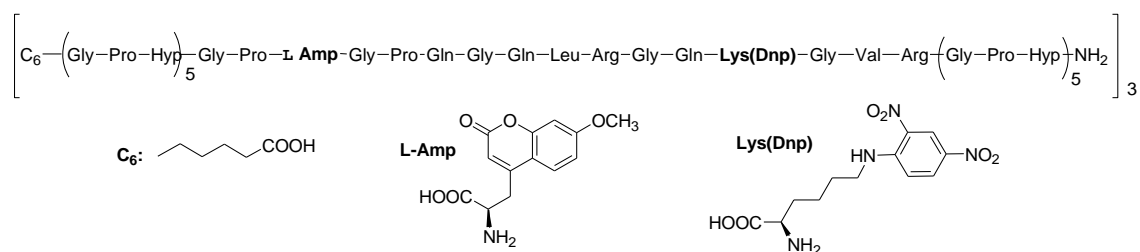
1.3.3 Modelos de Colágeno y Estabilidad de la Triple Hélice

La estabilidad de la triple hélice de colágeno ha sido principalmente estudiada en modelos de pequeño tamaño, formados por entre 5 y 20 tripletes (X-Y-Gly) por hebra.(84-85) Estos modelos sintéticos son mucho más fáciles de manipular y estudiar que el colágeno natural, y han permitido investigar distintos aspectos cinéticos y de estabilidad de la triple hélice,(86-87) analizar los efectos de mutaciones puntuales,(88) y caracterizar su estructura mediante difracción de rayos-X(89) y RMN.(90) La composición particular de estos modelos es muy variada, siendo la secuencia prototípica (Pro-Hyp-Gly)_n o bien modelos con una proporción importante de secuencia prototípica los más habitualmente considerados.

En esta Tesis se han analizado varios modelos de triple hélice de colágeno para los cuales se disponía de distinta información experimental relativa a su estructura, estabilidad o reactividad frente a peptidasas. En particular, hemos estudiado un modelo de secuencia puramente prototípica con 10 tripletes por hebra (Pro-Hyp-Gly)₁₀ que denominamos **POG10**.(89, 91) También se ha considerado un modelo de triple hélice con la secuencia (Pro-Hyp-Gly)₃-Ile-Thr-Gly-Ala-Arg-Gly-Leu-Ala-Gly-Pro-Hyp-Gly-(Pro-Hyp-Gly)₃. Este modelo, que denominamos **T3-785**.(81) presenta tres tripletes prototípicos en los extremos *N*- y *C*-terminal de cada hebra y una secuencia central que corresponde a un fragmento del colágeno tipo III. Otro sistema analizado ha sido el **THP-1**.(92) que es un modelo de la hebra α1 del colágeno tipo I con la secuencia correspondiente a la región α(1)772-786, que es reconocida e hidrolizada por las metaloproteinasas de la matriz (MMPs). Presenta 106 residuos por hebra, seis tripletes prototípicos en el extremo *N*-terminal y un enlace covalente entre las hebras en el extremo *C*-terminal según se muestra en el esquema:



donde *Ahx* es el ácido 6-aminohexanóico. Finalmente, el último modelo investigado es el denominado **fTHP-5**.(93) Se trata de un homotrímero cuya secuencia central es representativa de la zona de corte del colágeno III, e incorpora dos grupos fluorogénicos (*L-Amp* y *Lys(Dnp)*) que permiten el seguimiento cinético del proceso de hidrólisis de la triple hélice.



La estabilidad de estos y otros modelos de triple hélice ha sido caracterizada experimentalmente mediante técnicas calorimétricas y de dicroísmo circular. Desde un

punto de vista termodinámico, el proceso de formación de la triple hélice puede considerarse que ocurre en una sola etapa: (94)



donde TH es la triple hélice ya formada y H corresponde a la cadena o hebra aislada. La correspondiente constante de equilibrio K puede escribirse *aproximadamente* como:

$$K \approx \frac{[TH]}{[H]^3} = \frac{F}{3c_0^2(1-F)^3}, \quad (1.22)$$

siendo $c_0 = 3[TH] + [H]$ la concentración total de cadenas o hebras, y $F = 3[TH]/c_0$ el grado de helicidad de la muestra. Al calentar una muestra que contiene TH de cualquier modelo colagénico, el colágeno se disocia en sus hebras a medida que aumenta la temperatura. La temperatura a la cual $F = 0.5$, situación en la que la mitad de las hebras están asociadas en forma de triple hélice, se denomina *temperatura media de transición* (T_m) y es frecuentemente utilizada como medida de la estabilidad de la triple hélice correspondiente. La Figura 1.3 muestra como ejemplo las curvas de transición para las secuencias $(\text{ProProGly})_{10}$ y $(\text{ProHypGly})_{10}$, donde el valor de F a cada temperatura fue obtenido a partir de estudios de dicroísmo circular.(95)

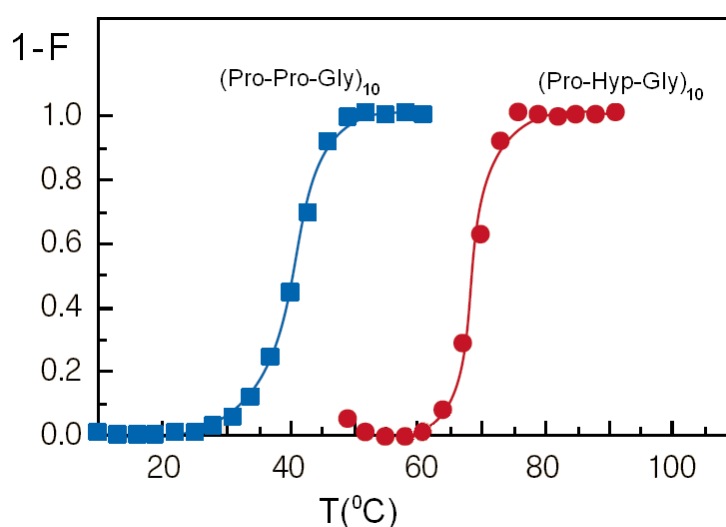


Figura 1.3: Ejemplo de curvas de transición. El experimento fue realizado con una concentración de las muestras de 0.23mM. Las temperaturas medias de transición son de 41 ± 1 y 69 ± 1 °C para $(\text{ProProGly})_{10}$ y $(\text{ProHypGly})_{10}$, respectivamente. La temperatura fue aumentada 3 °C cada 5 min.

Sabiendo que:

$$\Delta G^0 = -RT \ln K = \Delta H^0 - T\Delta S^0 \quad (1.23)$$

y combinando las ecuaciones (1.22) y (1.23) para $F = 0.5$ llegamos a:

$$T_m = \frac{\Delta H^0}{\Delta S^0 + R \ln(0.75c_0^2)} \quad (1.24)$$

En las ecuaciones anteriores, ΔG^0 es la energía libre de *Gibbs* estándar del proceso de desnaturalización, mientras que ΔH^0 y ΔS^0 son la entalpía y la entropía estándar, respectivamente. La ecuación (1.24) muestra la principal desventaja de utilizar T_m como medida de la estabilidad de la triple hélice, que no es otra que su dependencia con la concentración. No obstante, es frecuentemente utilizada para tal propósito y, de hecho, se han desarrollado métodos empíricos para predecir la temperatura media de transición en condiciones estándar a partir de la secuencia de aminoácidos de la triple hélice.(96)

Auxiliándose de datos de dicroísmo circular (CD), como los mostrados en la Figura 1.3, es posible obtener una entalpía de van't Hoff. Utilizando la ecuación de van't Hoff y la expresión de la constante de equilibrio (ecuación (1.22)), la entalpía puede ser evaluada a partir de la pendiente de la curva F vs T en la temperatura media de transición:

$$\Delta H_{vH}^0 = 8RT_m^2 \left(\frac{dF}{dT} \right)_{F=0.5} \quad (1.25)$$

Una vez determinada la entalpía y conocida también la constante de equilibrio K , el resto de las magnitudes termodinámicas, incluyendo ΔG^0 , pueden calcularse fácilmente a partir de las expresiones ya formuladas. Sin embargo, a pesar de que el protocolo anterior es el más utilizado para analizar la estabilidad de las hélices de colágeno, se ha comprobado que el modelo de una sola etapa y el posterior tratamiento con la ecuación de van't Hoff pueden tener grandes incertidumbres. De hecho, en la bibliografía

aparecen valores muy dispares para el mismo modelo de triple hélice, el denominado **POG10**.(87)

Una alternativa recomendada por algunos autores consiste en el empleo de la calorimetría diferencial de barrido (DSC).(87) Aunque tampoco está exenta de dificultades, estas pueden minimizarse disminuyendo en la medida de lo posible la velocidad de barrido. De este modo, Nishi y col. aseguran haber utilizado una velocidad de barrido lo suficientemente baja ($0.1 \text{ K}\cdot\text{min}^{-1}$) como para mantener la condición de equilibrio en todo momento. En esta Tesis nos hemos valido de los valores de ΔH , ΔS y ΔC_p reportados por estos autores a una temperatura de referencia $T^o = 344.9\text{K}$ para el modelo **POG10**. A partir de ellos, hemos estimado el valor de la variación de energía libre ΔG a la temperatura de interés $T = 298\text{K}$ utilizamos la expresión:

$$\Delta G(T) = \Delta H(T^o) + \Delta C_p(T - T^o) - T \left[\Delta S(T^o) + \Delta C_p \ln\left(\frac{T}{T^o}\right) \right]. \quad (1.26)$$

El valor finalmente obtenido para $\Delta G(T = 298\text{K})$ fue de -6.4kcal por mol de péptido (o cadena), es decir, correspondiente al equilibrio: $H \rightleftharpoons 1/3TH$.

La estabilidad de las triples hélices está directamente relacionada con la secuencia de aminoácidos de cada una de las tres hebras que la integran. En el colágeno natural se puede establecer una relación directa entre el contenido de imido-ácidos y la estabilidad de la triple hélice. Se sabe que, en general, a medida que la composición tiende hacia una secuencia prototípica del tipo (Pro-Hyp-Gly)_n, la estabilidad del sistema aumenta. Los anillos de prolina en posición **X** y de hidroxyprolina en posición **Y**, entre otros factores, estabilizan fuertemente aquellos valores de los ángulos Φ y Ψ (los ángulos diedros que determinan la conformación de la cadena principal) característicos de la triple hélice. En otras palabras, la presencia de los residuos Pro y Hyp estabiliza aquellos conformeros compatibles con las restricciones geométricas que impone la estructura en triple hélice.

Un análisis más detallado ha permitido comprender mejor por qué la hidroxilación de Prolinas en posición **Y** para formar la 4(R)-Hidroxiprolina (Hyp), proceso que ocurre después de la expresión de la proteína, es tan importante para la formación de las fibras

de colágeno(91, 97) y para la estabilidad de la triple hélice.(98) Inicialmente se pensó que la estabilidad extra que aporta la hidroxilación en posición **Y** se debía a la formación de puentes de hidrógeno mediados por moléculas de agua. De hecho, estos puentes de hidrógeno fueron observados mediante difracción de Rayos-X en colágeno cristalino.(80) Sin embargo, distintos estudios mostraron que la estabilidad relativa de las secuencias (Pro-Pro-Gly)_n y (Pro-Hyp-Gly)_n se mantiene en disolventes no acuosos, incluso después de una eliminación exhaustiva de las posibles trazas de agua.(99) Además, se comprobó que la sustitución de la 4(R)-Hidroxiprolina por 4(R)-Fluoroprolina (Flp) también aumenta la estabilidad térmica de la triple hélice.(95) Por tanto, parece improbable que estos puentes de hidrógeno mediados por agua contribuyan significativamente a la estabilidad del colágeno, probablemente debido al coste entrópico que supone su formación. Los estudios más recientes demuestran que el aumento de la estabilidad se debe, principalmente, al efecto inductivo del sustituyente hidroxilo en posición 4R y a su consiguiente influencia sobre el equilibrio conformacional del anillo de Prolina y del enlace peptídico.(100) Como se ha dicho, la estructura en triple hélice impone ciertas “restricciones” geométricas a los diedros del *backbone*, y estas restricciones son características de cada posición **X**, **Y** o **Z**. En la Figura 1.4 se resalta cómo la hidroxilación de la Prolina desplaza, por efecto inductivo, el equilibrio conformacional hacia el conformero donde el grupo -OH se coloca en posición axial, cuyos ángulos Φ y Ψ son justamente los apropiados para la posición **Y** de la triple hélice.

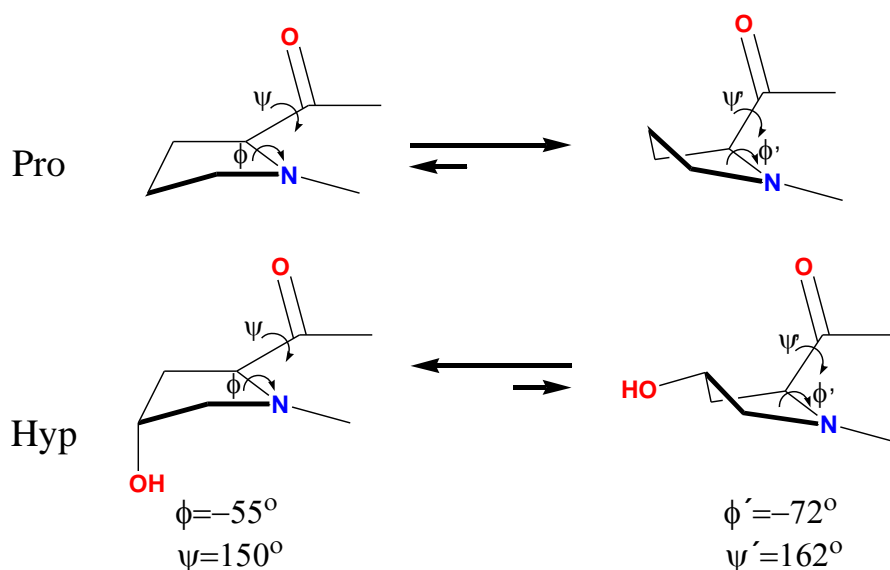


Figura 1.4: Desplazamiento del equilibrio conformacional al hidroxilar la posición **Y** en una triple hélice. Se muestran además los valores aproximados de los diedros del *backbone*.

1.3.4 Estudios Teóricos previos sobre Modelos de Colágeno

Debido al interés que tiene esta familia de proteínas, no sólo desde el punto de vista biológico sino también como posibles *biomateriales*,⁽¹⁰¹⁾ se han llevado a cabo numerosos estudios teóricos en modelos de colágeno más o menos realistas. El uso de la dinámica molecular (*Molecular Dynamics*, MD) ha permitido, por ejemplo, evaluar los efectos del disolvente sobre la estructura nativa del colágeno. De este modo, se ha comprobado que los campos de fuerza actualmente disponibles son capaces de reproducir la estructura de la triple hélice de colágeno y que ésta necesita del agua para mantener su estructura nativa.⁽¹⁰²⁾ Así, los resultados de la simulación MD de la secuencia (Pro-Hyp-Gly)₄-Pro-Hyp-Ala-(Pro-Hyp-Gly)₄ (campo de fuerzas parm94 y 1 ns de tiempo de simulación) mostraron un buen nivel de acuerdo con la información geométrica contenida en la correspondiente estructura de rayos-X (código PDB:1cgd).⁽¹⁰³⁾ No obstante, se observó una diferencia apreciable en los valores de RMSD (*Root Mean Square Deviation*) calculados para los átomos C α en toda la molécula o únicamente en la zona central, que ha sido racionalizada en nuestro trabajo posterior.

Basándose en los datos disponibles para compuestos modelo, se ha afirmado que las regiones de colágeno con un bajo contenido en imido-ácidos son susceptibles de experimentar cambios conformacionales en los que la triple hélice presenta una estructura más extendida.⁽¹⁰⁴⁾ Esta hipótesis se ha investigado computacionalmente para el modelo **T3-785**. Mediante simulaciones de dinámica molecular con restricciones harmónicas fue posible estimar la diferencia de energía libre entre un estado inicial próximo a la estructura de rayos-X y un segundo estado en el que la triple hélice estaba ligeramente desenrollada en su parte central. Ambos estados resultaron ser cuasi-energéticos, estando separados por una pequeña barrera energética de tan sólo 2.0 kcal/mol. Sin embargo, en estas simulaciones sólo se añadieron tres capas de disolvente explícito alrededor de la triple hélice, y no se calibró la bondad del campo de fuerzas utilizado (CHARMM(1)) para reproducir la estructura y propiedades de los sistemas colagénicos.

Las zonas de corte por las que son hidrolizadas las moléculas de colágeno suelen tener un bajo contenido en imido-ácidos y las mutaciones en esta zona pueden tener consecuencias para la salud.(105) Se ha estudiado mediante dinámica molecular cómo esas mutaciones en el entorno de la zona de corte afectan geoméricamente a la estructura en triple hélice.(105) El proceso de propagación en la formación de la triple hélice, que ocurre desde el extremo C- al N-terminal, ha sido también estudiado en modelos prototípicos(106) con el objeto de analizar el efecto de una mutación Gly→Ser del tipo de las encontradas en distintas variantes de la enfermedad osteogénesis imperfecta. Después de construir un modelo parcialmente desenrollado, y mediante una dinámica molecular en la que se imponen restricciones geométricas que facilitan cinéticamente el *folding* o el proceso contrario dependiendo del objetivo a seguir (el nombre de esta técnica es BMD, (107) *Biased Molecular Dynamics*), se ha encontrado que la mutación imposibilita la propagación del *folding*.

Las moléculas de colágeno pueden ser homotrímeros, si la secuencia de las tres cadenas es la misma, o heterotrímeros en caso contrario. Adicionalmente, en un colágeno heterotrímero, cada molécula puede tener más de un *registro*, es decir, distintos desplazamientos relativos entre las diferentes cadenas. El colágeno tipo IV de composición $[\alpha 1(\text{IV})]_2\alpha 2(\text{IV})$, por ejemplo, podría formar en principio tres heterotrímeros distintos con los registros $\alpha 1\alpha 1\alpha 2$, $\alpha 1\alpha 2\alpha 1$ y $\alpha 2\alpha 1\alpha 1$. Simulaciones de dinámica molecular de estos dos últimos han desvelado la importancia del registro en la estabilidad y en el *folding* de la triple hélice(108). No sólo se ha estudiado la asociación de hebras para formar una triple hélice, sino también la asociación de las propias hélices entre sí para formar una estructura supramolecular.

Recientemente, se han hecho simulaciones de dinámica molecular sobre un modelo relativamente reducido de colágeno I con 45 residuos por cadena. En este caso se realizaron cálculos de desnaturalización (*unfolding*) local de cada cadena del modelo, concluyendo que este proceso de desnaturalización parcial es más favorable que ocurra en la cadena $\alpha 2(\text{I})$ que en la $\alpha 1(\text{I})$.(109) Sin embargo, los autores descartaron de antemano el registro $\alpha 1\alpha 2\alpha 1$ que es el registro sugerido experimentalmente.(86, 110) La justificación utilizada para ello fue simplemente que sólo el registro $\alpha 1\alpha 1\alpha 2$ se mantenía en una estructura estable de triple hélice.

Los estudios QM del colágeno se han limitado en su mayoría a modelos de pequeño tamaño en los que se ha analizado el efecto de mutaciones en las glicinas sobre la estabilidad de un triplete particular.(111) En estos estudios, se ha calculado la estabilidad relativa de pequeños modelos de triple hélice, observando que ésta varía según la referencia energética que se tome, ya sean las hebras separadas, las hebras separadas y relajadas o los aminoácidos por separado. También se ha estudiado, mediante cálculos QM, el equilibrio conformacional del anillo de la Prolina y de la Hidroxiprolina, debido al importante papel que juegan ambos residuos en la estabilidad del colágeno. Los resultados teóricos indican que los distintos conformeros optimizados en ambos casos son geoméricamente casi idénticos,(112) pero que hay un cambio en la preferencia energética del *puckering* del anillo.(113) Precisamente, la parametrización del *puckering* del anillo y de las cargas de la Hidroxiprolina es una etapa previa esencial para las posteriores simulaciones de dinámica molecular del colágeno, y los cálculos teóricos sobre modelos peptídicos han sido la principal herramienta de parametrización.(114)

1.4 Cálculos de Energía y Entropía en Modelos de Colágeno: Desafíos Metodológicos

La idea central de esta Tesis es la introducción de nuevas propuestas para una mejor estimación de la entropía y la energía de sistemas biomoleculares, y su aplicación durante el estudio de la triple hélice de colágeno. A este respecto, la caracterización de la estructura y de la estabilidad de modelos de colágeno pasa por la realización de simulaciones de dinámica molecular utilizando potenciales clásicos, con el objetivo de obtener información estructural lo más completa posible tanto del estado de triple hélice como de los estados disociados. Y en este caso, cuando hablamos de información estructural, nos referimos a los dos estados (triple hélice y desnaturalizado) como colectivos estadísticos. Como hemos mencionado anteriormente, el disolvente juega un papel importante en la estabilidad de los modelos de colágeno. (81, 115) De hecho, debido a su estructura lineal, las triples hélices cuentan con una gran superficie accesible al disolvente. En este trabajo se ha prestado una especial atención a este punto, realizando las simulaciones en disolvente explícito, ya que se ha comprobado las

deficiencias y grandes diferencias que puede haber al optar por un modelo continuo de disolvente.(116) Además, las simulaciones se han realizado utilizando condiciones periódicas y correcciones de largo alcance para las contribuciones electrostáticas.

Con independencia de la metodología de simulación empleada, en sistemas tan complejos como las biomoléculas no dispondremos en general de un muestreo del espacio configuracional, y mucho menos del espacio de fases, lo suficientemente representativo. El trabajo se centra pues, en hacer la mejor estimación posible de las magnitudes de interés con la “reducida” información estructural disponible. La magnitud más sensible a la no completitud del muestreo es, por supuesto, la entropía. De hecho, en el proceso de formación del colágeno la entropía juega un papel si cabe más importante, ya que se pasa de una estructura disociada muy flexible a un estado de triple hélice con grandes restricciones conformacionales.(26) Incluso la propia cinética de formación del colágeno está controlada en gran medida entrópicamente.(25) La correcta estimación de la variación de entropía en procesos biomoleculares, que es uno de los retos de la actual Química Computacional, es de vital importancia para este trabajo. Sin embargo, el clásico análisis de los modos normales vibración no nos permite capturar el efecto de este cambio de flexibilidad, ya que cada cálculo sólo captura la incertidumbre asociada a un solo mínimo de la superficie de energía potencial. Necesitamos por tanto, dar cuenta también de la entropía asociada a la ocupación de diferentes mínimos dentro de la superficie,(3, 17, 51) lo que nos lleva a desarrollar nueva metodología en esta temática de gran actualidad.

Por otra parte, a pesar de que estamos aún lejos de hacer simulaciones del comportamiento dinámico de grandes biomoléculas por métodos puramente QM, se han dado pasos importantes para estimar las energías QM de grandes sistemas a partir de las de sus fragmentos. A diferencia de los cálculos de entropía por métodos estadísticos, que necesitan una enorme cantidad de muestreo, se pueden hacer buenas aproximaciones de la energía media de un sistema macromolecular con un conjunto bastante más reducido de estructuras representativas. Es en este punto donde los cálculos QM aproximados de grandes estructuras podrían mejorar las estimaciones de energéticas que se realizan rutinariamente con los campos de fuerzas empíricos, mediante re-cálculos sobre las estructuras obtenidas mediante mecánica molecular. Nuestra propuesta denominada TFEM (*Thermodynamic Fragment Energy Method*) está

optimizada para estructuras lineales como el colágeno, aunque es también aplicable a estructuras más compactas con un coste mayor.

Para analizar la estabilidad global de una biomolécula en disolución acuosa, es evidente que además de la energía y la entropía del soluto, son también importantes otras componentes adicionales de energía libre que no han sido analizadas en esta Tesis. La energía de solvatación, por ejemplo, es una magnitud extraordinariamente difícil de evaluar con precisión, ya que los modelos desarrollados hasta la fecha son muy simples e incompletos. Así la energía de cavitación, que es la energía libre asociada a la formación de una cavidad con la forma de nuestro soluto, se calcula rutinariamente como el producto de la tensión superficial del disolvente por el área de la superficie molecular.⁽²³⁾ Sin embargo, esta enorme simplificación de un proceso tan complejo hace que muchas veces los resultados no concuerde con lo esperado.⁽¹¹⁷⁻¹¹⁸⁾ Por este y otros motivos, no es posible afirmar que el estudio de la estabilidad de los distintos modelos de colágeno sea uno de los objetivos de esta Tesis, pero sí lo es el análisis de las contribuciones entrópicas del soluto a dicha estabilidad.

Objetivos

La presente Tesis tiene como objetivo principal introducir mejoras en el cálculo de la energía y la entropía en sistemas biomoleculares, siendo los modelos de colágeno el principal objeto de estudio. Con esta consideración global en mente, enunciamos brevemente los siguientes objetivos específicos de esta Tesis Doctoral:

- *Diseño, implementación y validación de algoritmos eficientes basados en criterios de cutoff para estimar la entropía conformacional de una molécula de soluto (una biomolécula) a partir de simulaciones moleculares.* Al plantear este objetivo se está asumiendo la descomposición de la entropía total (excluyendo los grados de libertad de rotación y traslación) en componentes vibracional y conformacional. No obstante, un sub-objetivo será evaluar la bondad de dicha partición en sistemas de pequeño tamaño cuya entropía absoluta haya sido determinada experimentalmente. Al desarrollar técnicas basadas en criterios de *cutoff* para calcular la entropía conformacional, se pretende evitar los problemas de convergencia de los métodos MIE en sistemas de gran tamaño. Pero es también igualmente necesario eliminar la posible ambigüedad en la selección del *cutoff*, buscando alguna estrategia para seleccionar el valor más adecuado para cada caso. Por supuesto, el cumplimiento global de este objetivo requiere la implementación en forma de código de programación de los sucesivos métodos diseñados en esta Tesis para calcular las entropías conformacionales.
- *Diseño de un método para estimar la energía QM de biomoléculas a partir de las energías de sus fragmentos convenientemente definidos.* Como vimos en el apartado 1.2, el Método MBE establece el marco general para este tipo de aproximaciones basadas en fragmentos. Aprovecharemos el hecho de que las expansiones MBE contienen muchos más términos de los necesarios para obtener valores de energía con una precisión razonable. Teniendo en cuenta que la energía es una función de estado, nos podemos plantear una reacción formal de fragmentación de toda la molécula acoplada a un criterio de *cutoff*, lo que

nos permitiría hacer estimaciones de la energía total. Por su naturaleza, este planteamiento facilitaría la utilización de metodologías ya establecidas para la inclusión de los efectos del disolvente mediante los distintos modelos disponibles, o incluso combinarlo con metodologías de tipo QM/MM.

- *Caracterización estructural y dinámica de modelos de colágeno mediante simulaciones de dinámica molecular en disolución acuosa.* Se analizarán aspectos tales como la estabilidad dinámica de los principales contactos presentes en la estructura en triple hélice, la interacción con el disolvente, la distorsión de la estructura promedio, la naturaleza de los estados desnaturalizados, etc. Además de complementar la información experimental previa, obtenida mediante estudios de difracción de rayos X y RMN principalmente, la comparación simulación/experimento nos permitirá validar la metodología utilizada, principalmente el campo de fuerzas seleccionado y la parametrización desarrollada para los residuos no estándar.
- *Analizar la influencia de los efectos entrópicos sobre la estabilidad de los modelos colágeno.* Es importante cuantificar estos efectos debido a que es esperable que sean significativos en la formación de una de la triple hélice, ya que en su formación, pasa de un estado desestructurado a una estructura esencialmente rígida. Veremos cómo estos modelos de colágeno son unos sistemas ideales en los que estudiar los cambios en la entropía conformacional.

En su conjunto, los resultados de este trabajo nos permitirán avanzar hacia la mejora de los protocolos para predecir la estabilidad de biomoléculas, en particular en sistemas colagénicos.

Capítulo II

Discusión de Resultados y Publicaciones

Los resultados serán presentados en este segundo capítulo de la memoria dentro de la modalidad abreviada de compendio de publicaciones. Para ello, tal y como establece la actual normativa de los estudios de tercer ciclo de la Universidad de Oviedo, se incluirán las diferentes publicaciones agrupadas dentro un marco común para una mayor comprensión de la unidad temática de la Tesis. De este modo, las publicaciones no serán presentadas en un orden cronológico estricto, sino que se seguirá un orden temático inicial manteniendo después la cronología dentro de cada tema.

El primer tema, y al que más tiempo y esfuerzo hemos dedicado, se centra en el desarrollo de metodología para la estimación de entropías a partir de simulaciones de dinámica molecular. Concretamente se calcula la parte conformacional de la entropía de distintos sistemas (macromoléculas, péptidos e hidrocarburos) con varios desarrollos metodológicos. Dentro de esta temática incluimos un total de 6 artículos de investigación, de los cuales 4 ya han sido publicados en revistas internacionales (*J. Phys. Chem. B*, *J. Chem. Theor. Comput.*, *Entropy* and *Proteins*) mientras que los últimos dos trabajos se encuentran en periodo de revisión. Además, los resultados mostrados en los artículos y manuscritos se complementan con cálculos adicionales de entropía conformacional, que aplican la última propuesta metodológica desarrollada en esta Tesis, la denominada CC-MLA (*Correlation Corrected Multibody Local Approximation*), a toda la serie de modelos de colágeno considerados en el conjunto de nuestra investigación (POG10, T3-785, THP-1 y fTHP-5).

El segundo gran tema incluye básicamente una propuesta metodológica sobre cómo estimar con una precisión razonable, la energía QM de un sistema de gran tamaño como la triple hélice de colágeno, a partir de las energías de fragmentos previamente definidos. Esta segunda temática de la Tesis ha dado lugar a la publicación de un artículo de investigación (*J. Chem. Theor. Comput.*). Además, cálculos de energía más precisos que los inicialmente publicados fueron presentados como publicación asociada al congreso *International Conference on Computational and Mathematical Methods in Science and Engineering*.

Seguidamente, resumimos los principales resultados obtenidos dentro de cada tema. El objetivo es proporcionar una visión global de los avances metodológicos o en la caracterización de la triple hélice de colágeno, intentando conectar los distintos trabajos entre sí. Los detalles más técnicos expresados en términos de ecuaciones o de resultados numéricos aparecen recogidos en los distintos artículos y manuscritos, para los que se adjunta también la información suplementaria enviada a la revista.

2.1 Cálculos de Entropía en Biomoléculas a partir de Dinámica

Molecular

En nuestro primer trabajo “*Entropic Control of the Relative Stability of Triple-helical Collagen Peptide Models*”, se desarrolla la parametrización necesaria para simular, en el contexto del campo de fuerza AMBER03, el residuo no estándar 4R-hidroxi prolina incluido en los modelos POG10 y T3-785 presentados en el capítulo anterior. Los análisis estructurales realizados permiten comprobar la bondad de la aproximación MM, ya que se observa un buen acuerdo entre los resultados de las simulaciones y los datos estructurales obtenidos a partir de experimentos de RMN y difracción de rayos-X. Sin embargo, nuestras simulaciones también aportan nueva información sobre el comportamiento dinámico de la triple hélice o sobre la estructura del disolvente alrededor de la misma.

Como objetivo adicional, nos planteamos predecir la estabilidad de ambos modelos colagénicos POG10 y T3-785, estudiando en cada caso la formación de la triple hélice a partir de sus hebras aisladas. Observamos que mientras las triples hélices presentan

una cadena principal básicamente rígida, las hebras aisladas muestran una mayor flexibilidad aunque muy condicionada por el contenido en imido-ácidos de la secuencia. En el estado disociado, el modelo POG10 que tiene el máximo contenido posible de imido-ácidos (secuencia prototípica) es mucho menos flexible que el modelo T3-785. Este efecto se recoge en el cambio en la entropía conformacional asociada a la formación de la triple hélice y permite explicar la distinta estabilidad de ambos modelos de colágeno. Se inicia así nuestra aproximación al cálculo de entropías.

En este trabajo inicial, por primera vez se comprueba lo determinante que resulta el efecto entrópico conformacional de los anillos de Prolina y 4R-Hydroxiprolina sobre las estabilidades relativas de modelos de triple hélice. Para el cálculo de entropía, dividimos la entropía total excluyendo la traslación y la rotación, en vibracional y conformacional, es decir $S_{tot} = \langle S_{vib} \rangle + S_{conf}$, donde $\langle S_{vib} \rangle = \sum_i p_i S_{vib,i}$ es la entropía vibracional promediada sobre los diferentes mínimos de la superficie de energía potencial. Estos mínimos se obtienen optimizando un conjunto representativo de puntos de la dinámica utilizando el modelo de disolvente continuo GBSA.(119) Por otro lado, $S_{conform}$ es la entropía asociada a la incertidumbre de pertenecer a uno u otro mínimo (confórmero). Se calcula utilizando todos los puntos obtenidos de la simulación, después de transformar las densidades marginales asociadas a los ángulos diedros en estados conformacionales.

Esta partición de la entropía total será una constante en todos nuestros trabajos, ya que simplifica muchísimo la estimación de entropías en biomoléculas. Para superar los problemas de sesgo intrínsecos a la estimación estadística, se utiliza una aproximación MIE hasta orden 4 para el cálculo de la entropía conformacional utilizando sólo los diedros del *backbone*. Como veremos, este protocolo será mejorado sustancialmente más adelante con nuevas propuestas metodológicas. El cambio de entropía para la formación de la triple hélice, se combina entonces con el resto de contribuciones energéticas al correspondiente ΔG obtenidas mediante el protocolo MM-PBSA (*Molecular Mechanics Poisson-Boltzmann Surface Area*)(23). Se obtienen valores relativos y absolutos de ΔG que están en buen acuerdo con los datos experimentales reportados en la literatura.

Tenemos, pues, una herramienta que nos permite estimar la contribución de los cambios en la entropía conformacional en distintos procesos de interés. Más allá del estudio de la triple hélice, este avance metodológico puede tener más aplicaciones como se muestra en el segundo trabajo titulado “*Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and QM/MM Calculations*”, donde se incluyó el efecto de la entropía conformacional en la estimación de la energía libre de *binding* relativa de sendos péptidos unidos al centro activo de la enzima MMP-2. Para reproducir la estabilidad relativa de los dos complejos de Michaelis se utilizó el mismo protocolo MM-PBSA presentado en el trabajo anterior complementado con las estimaciones de la entropía conformacional basadas en un MIE de cuarto orden aplicado únicamente a las torsiones del *backbone* de los substratos peptídicos. Finalmente, los resultados obtenidos mostraron un buen acuerdo con los datos experimentales de K_M disponibles.

El tercer trabajo incluido en esta Tesis lleva por título “*Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations*”. Este es un trabajo exclusivamente metodológico recientemente publicado, que se elabora en un momento donde ya se habían logrado importantes avances tanto en la teoría, como en los algoritmos y en la eficiencia del programa que habíamos diseñado para el cálculo de las entropías conformacionales. Desde el punto de vista teórico, se logra reformular la expansión MIE reduciendo al mínimo su redundancia y evitando almacenar en memoria RAM la gran cantidad de entropías de los subconjuntos generados. Además, se propone un método riguroso para el proceso de discretización de las torsiones, mediante el que podemos obtener densidades de probabilidad marginales analíticas. Esto nos permite calcular sin dificultad la primera y segundas derivadas analíticas e identificar automáticamente los mínimos de las densidades de probabilidad marginales (paso esencial en la discretización). Todo esto, unido a un gran número de pequeñas mejoras en la eficiencia del código y a su implementación en paralelo, nos permitió ser más ambiciosos en los órdenes de la expansión MIE que podían ser analizados.

En el tercer trabajo también nos planteamos como objetivo comparar nuestros resultados teóricos con entropías experimentales. Para ello se seleccionaron pequeños alcanos (6-7 átomos de carbono) para los que se conoce su entropía experimental en

fase gas en condiciones estándar. Se obtuvo un buen acuerdo teoría/experimento al añadir las correcciones conformacionales a las entropías promedio RRHO (*Rigid Rotor Harmonic Oscillator*). Las entropías RRHO se obtuvieron con el funcional híbrido B3LYP y la base de Dunning cc-pVTZ (*Correlation Consistent Polarized Valence Triple Zeta*) sobre un conjunto de estructuras extraídas de dinámica molecular. Este nivel de cálculo fue seleccionado porque reproduce muy bien valores de frecuencias vibracionales y de entropías experimentales en los casos en que existe un solo mínimo conformacional.(120)

Como último análisis incluido en este tercer trabajo, se calculó la entropía conformacional de un conjunto de pequeños péptidos (cinco dipéptidos y un hexapéptido). Estos son sistemas mayores que los alcanos, donde los efectos de orden superior van a ser más importantes. Pero a la vez son sistemas relativamente pequeños, cuyo análisis no es demasiado costoso computacionalmente. Gracias a ello pudimos comprender un poco más el comportamiento en el orden de la expansión MIE. Comprobamos, por ejemplo, lo costoso que resulta la convergencia en el orden de esta aproximación a medida que aumenta el tamaño del sistema. La necesidad de un método que supere estas limitaciones nos lleva a nuestro próximo desarrollo.

El cuarto trabajo “*Distinguishability in Entropy Calculations: Chemical Reactions, Conformational and Residual Entropy*” complementa el trabajo anterior (y cualquier otro donde se estimen entropías conformacionales o configuracionales), en un tema en el que fácilmente se puede cometer errores, las “correcciones” a la entropía debidas a indistinguibilidad átomos o moléculas idénticas. Mediante el análisis de diferentes ejemplos y el uso de conceptos como la entropía conformacional y residual, se muestra en este trabajo que las entropías calorimétricas experimentales pueden ser reproducidas teóricamente tanto desde un formalismo donde los átomos idénticos son indistinguibles, como desde un punto de vista clásico donde todos los átomos son distinguibles. Veremos que la corrección ampliamente utilizada en los cálculos de entropía debido al número de simetría y la indistinguibilidad de partículas no es obligatoria, como corrección a posteriori, para obtener valores precisos de entropías absolutas y relativas. Más importante aún, es que se demuestra que para *cualquier* reacción química de *cualquier* naturaleza, considerar las partículas como distinguibles es igual de válido

siempre que actuemos de manera coherente en el cálculo de todas las contribuciones entrópicas.

En el quinto trabajo titulado “*Multibody Local Approximation for Conformational Entropy Calculations on Biomolecules*” obtenemos uno de los resultados metodológicos más importantes de la Tesis. Las expansiones de naturaleza MBE o MIE, que son básicamente equivalentes, son muy utilizadas para estudiar propiedades de grandes sistemas, teniendo como ventaja principal su generalidad. Es decir, son exactas para cualquier sistema y cualquier propiedad si extendemos la expansión hasta el máximo orden posible (el número de elementos, variables, o subsistemas en consideración). Sin embargo, existen dos importantes inconvenientes. Por un lado, lo costoso de la inclusión de términos de órdenes superiores debido a la naturaleza combinatoria de la expansión. Por otro lado está el hecho, muchas veces ignorado, de que el error de truncamiento de la expansión no decrece necesariamente con el orden.

En este trabajo aprovechamos nuevamente la redundancia de estas expansiones para obtener lo que llamamos AMIE (*Approximate MIE*), que resulta ser una buena aproximación de la expansión original cuando nos planteamos ésta última como dependiente de un criterio de *cutoff*. Se comprueba que a partir de AMIE se obtiene una expresión excepcionalmente eficiente que captura, sin calcularla explícitamente, la correlación a todos los órdenes compatible con un *cutoff* determinado. Es decir, se transforma una expresión que es dependiente del orden a otra dependiente del *cutoff*. Se conserva así la generalidad de la expresión original, a la vez que se superan sus deficiencias. Se propone además un criterio para seleccionar el *cutoff* más adecuado dadas la naturaleza de nuestro sistema y la extensión de nuestro muestreo mediante la dinámica molecular. Finalmente, utilizando la dinámica molecular de una de las hebras del modelo fTHP-5, se comparan nuestros resultados con la aproximación QHA después de añadirle a la componente conformacional el resto de componentes entrópicas. De esta forma se verifica que somos capaces de obtener una cota superior de la entropía total mucho más ajustada que la proporcionada por el ampliamente extendido método QHA.

El último trabajo dentro de este primer tema dedicado a la entropía lleva por título “*CENCALC: A New Program for Conformational Entropy Calculation of*

Macromolecules from Molecular Simulations". Corresponde a la formalización como código de programación de todos los métodos que hemos desarrollado para el cálculo de las entropías conformacionales. El paquete CENCALC, que se distribuirá libremente bajo licencia GNU, está compuesto por tres programas *cencal_omp.f90*, *cencalc_prep.f90* y *get_tor.py* cuyas características y funciones específicas se describen en el manuscrito. En éste se presenta de modo muy resumido toda la teoría de los métodos implementados y se aplican al análisis de la dinámica de un péptido de 10 residuos que llamamos GNR, discutiendo además la eficiencia de los distintos métodos. Como ejemplo ilustrativo adicional, estudiamos el cambio de entropía conformacional del GNR al unirse a la metaloproteinasa de la matriz MMP-7. Con este manuscrito, se adjunta el manual del programa y, por otro lado, el código y ejemplos en formato digital.

2.1.1 Compendio de Publicaciones

En esta sección se recopilan las publicaciones y manuscritos que abordan el primer tema de la Tesis. En ellos se pueden encontrar en detalle los resultados de los estudios introducidos anteriormente así como todo el material suplementario.

**2.1.1.1 *Entropic Control of the Relative Stability of Triple-helical
Collagen Peptide Models***

Ernesto Suárez, Natalia Díaz and Dimas Suárez

J. Phys. Chem. B 2008, 112: 15248-15255

Entropic Control of the Relative Stability of Triple-helical Collagen Peptide Models

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

Departamento de Química Física y Analítica, Universidad de Oviedo,
C/ Julián Clavería, 8. 33006, Oviedo, Spain

Received: August 20, 2008

Herein, we show that current methodologies in atomistic simulations can yield reliable standard free energy values in aqueous solution for the transition from the dissociated monomeric form to the triple-helix state of collagen model peptides. The calculations are performed on a prototypical highly stable triple-helical peptide, [(Pro-Hyp-Gly)₁₀]₃ (POG10), and on the so-called T3–785 triple-helix mimicking a fragment from the type III human collagen, which is more thermally labile. On the basis of extensive MD simulations in explicit solvent followed by molecular-mechanical and electrostatic Poisson–Boltzmann calculations complemented with an accurate estimation of the nonpolar contributions to solvation, the computed free energy change for the aggregation processes of the POG10 and T3–785 peptides leading to their triple-helices is –6.6 and –6.1 kcal/mol, respectively. For POG10, this value is in agreement with differential scanning calorimetric data. However, it is shown that conformational entropy, which is estimated by means of an expansion of mutual information functions, preferentially destabilizes the triple-helical state of T3–785 by around 4.6 kcal/mol, thus explaining its lower thermal stability. Altogether, our computational results allow us to ascertain, for the first time, the actual thermodynamic forces controlling the absolute and relative stability of collagen model peptides.

Introduction

Collagen is the most abundant extra-cellular protein in animals, with a wide variety of physiological roles and involvement in pathological processes. The molecule presents a characteristic triple-helix structure^{1,2} composed of three peptide chains, each in an extended, left-handed polyproline II-like helix, which are staggered by one residue and then supercoiled about a common axis in a right-handed manner. Triple-helices are characterized by repetitions of the triplet X–Y–Gly pattern, which extend about 1000 residues long in the natural collagen molecules. Staggered arrays of collagen molecules form fibrils, which arrange to form collagen fibers with superior mechanical properties.³

Important milestones on the collagen structure and stability have been achieved during the last few years. The spatial arrangement of type I collagen molecules in fibrils has been determined by X-ray crystallography.⁴ It has also been shown that the energetically stable conformation of type I collagen under physiological conditions is a random coil rather than a triple-helix.⁵ However, natural collagen is still very difficult to handle in the laboratory and, therefore, to investigate the thermal stability and folding of the triple-helix domain, many synthetic triple-helical peptides (THPs) with 30–45 amino acids per chain have been characterized using a wide array of experimental techniques.^{6,7} In general, the stability of a THP is expressed in terms of the midpoint temperature (T_m) for the transition from the triple-helical complex to the dissociated monomers as obtained from circular dichroism spectroscopy profiles. However, thermodynamic parameters derived from T_m data for the same THP can vary substantially, even as much as 20 kcal/mol, depending on the experimental conditions.⁸ Differential scanning calorimetric (DSC) measurements can be problematic

as well, owing to strong scanning rate dependency. These uncertainties in the THP thermodynamic parameters point to several difficulties in achieving equilibrium and unknown concentration dependencies.⁶

Clearly, a detailed knowledge of the structure and free energy in solution of THPs is required to better understand the stability of the collagen triple-helix. In addition, such knowledge could be useful for the design of biofunctional synthetic THPs.^{9,10} To this end, we report here the results of a well-balanced computational protocol aimed at the evaluation of the relative stability of THPs. The calculations were performed on a prototypical THP molecule, [(Pro-Hyp-Gly)₁₀]₃ (POG10), which has the highest possible imidic acid content. The Pro and 4(*R*)-hydroxyproline (Hyp) residues stabilize preferentially the backbone conformations that are compatible with the geometrical restrictions imposed by the triple-helical symmetry. Thus, POG10 turns out to be quite stable, with T_m values around ~60 °C.¹¹ In addition, we also investigated the T3–785 triple-helix, [(Pro-Hyp-Gly)₃(Ile-Thr-Gly)-(Ala-Arg-Gly)-(Leu-Ala-Gly)-(Pro-Hyp-Gly)₄]₃, whose central region mimics a fragment from the type III human collagen. Replacement of imidic acids by other amino acids reduces the thermal stability of the THP molecules, with the T_m value for T3–785 being 25 °C.¹¹ Prior computational studies have investigated some factors influencing the triple-helix conformation of these and other THPs.^{12–16} In this work, however, we performed molecular dynamics simulations in explicit solvent considering both the triple-helix and monomer states of the selected THPs. Subsequently, the standard free energy of the systems was computed using a refined version of the molecular-mechanical Poisson–Boltzmann approach complemented with an estimation of the conformational entropy based on the expansion of mutual information functions. Altogether, the computational results give new insight into the thermodynamic stability of the THPs.

* To whom correspondence should be addressed. Phone: +34-985103689; fax: +34-985103125; e-mail: dimas@uniovi.es.

Experimental Methods

MD Simulations of the Triple-helical Structures. The AMBER03 force field¹⁷ was used to model the systems. A set of atomic charges and specific torsions for the Hyp residue were derived following the prescriptions for parameter derivation as described in the AMBER03 protocol. Further details are given in the Supporting Information (Figure S1).

Starting coordinates for the triple-helical models were obtained from crystal structures 1CAG² (POG10) and 1BKV¹⁸ (T3–785). Both X-ray structures contain three peptide chains with 30 amino acids per chain. In the 1CAG structure, the three chains present an Ala residue in their central region in the place that should occupy a Gly. During molecular edition, this mutation was reversed back to the prototypical sequence to build up the POG10 structure. The protonation states for the ionizable residues were set to their normal ionization state at pH 7. The protein atoms, as well as all the water molecules of the crystal structure, were surrounded by a periodic box of water molecules that extended 15 Å from the protein atoms. This resulted in a box size of approximately 50 × 50 × 120 Å with about 9.500 water molecules. For the T3–785 system, three Cl[−] counterions were included to neutralize the system.

MD simulations were carried out using the PMEMD module of the Amber 9 package.¹⁹ The solvent molecules and counterions were initially relaxed by means of energy minimizations and 50 ps of MD. Subsequently, the full systems were minimized to remove bad contacts in the initial geometry and heated gradually to 300 K during 40 ps of MD. During the thermalization of the systems, harmonic constraints were imposed on the position of the protein atoms, but these were successively lowered and finally removed. Subsequently, the systems were coupled to a thermal and a hydrostatic bath at $T = 300$ K and $P = 1.0$ atm. The time step of integration was 2.0 fs and the SHAKE procedure on the X–H bonds was applied. The PME approach was used for nonbonded interactions. For the POG10 and T3–785 models, 20 ns trajectories were computed, and coordinates were saved for analysis every 500 time steps.

Principal component analysis (PCA) was performed as described by Sherer et al. in their MD study of DNA.²⁰ From the last 15 ns of each trajectory, the positional covariance matrix of the C α atoms was calculated using 1500 snapshots. To avoid end effects, the N- and C-terminal [(X–Y–Gly)₃] units were removed from the analysis. In all cases, the corresponding PCA eigenvectors of this matrix describe the largest part of the structural variance of the THP backbone atoms along the trajectories. To ease the interpretation of the deformations associated with the major PCA components, we followed the procedure described in detail by Sherer et al.²⁰ Thus, the corresponding PCA eigenvectors were projected onto the coordinates of selected MD snapshots, producing a set of transformed coordinates in which all elements, save that of the eigenvector of interest, were replaced with time-averaged values. Subsequently, the projection was reversed, and the resulting structures were visually inspected.

Monomer Peptides: MD Simulations and Conformational Search Calculations. Starting coordinates for the isolated chains of the POG10 and T3–785 peptides were taken from the first chain in the last snapshot of the triple-helix simulations. MD simulations were performed using the settings as described above. Conformational search calculations were performed using the LMOD program linked to the Amber package.¹⁹ LMOD implements a conformational search algorithm based on eigenvector following of low-frequency vibrational modes that allows

flexible docking and protein loop optimization.²¹ We employed the AMBER03 force field coupled with the Hawkins–Cramer–Truhlar pairwise Generalized-Born (GB) model.²² A total of 6000 LMOD iterations were computed by exploring 3 low-frequency vibrational modes. Eigenvectors were recalculated every 25 LMOD iterations.

500 snapshots from each MD trajectory were clustered using the MMTSB-tools.²³ The mutual similarity algorithm was employed by selecting a fixed cluster radius (90°) and considering only the backbone torsion angles. The structure in each cluster with the lowest deviation is taken as the cluster representative.

Molecular Mechanical and Poisson–Boltzmann Calculations. We estimated various free energy components of the solute molecules by performing MM-PB calculations^{24,25} on 500 snapshots extracted from the production phase of the different MD simulations, except for the monomer state of T3–785, for which 1000 snapshots were used. The snapshots were postprocessed through the removal of all solvent or counterions. The average MM-PB free energy of the set of structures was computed according to the following equation:

$$\bar{G}_{\text{MM-PB}} = \bar{E}_{\text{MM}} + 3RT + \bar{G}_{\text{PB-elec}} + \bar{H}_{\text{vdW solute-solvent}} - T\bar{S}_{\text{MM-GBSA}}^{\text{norm}} \quad (1)$$

where \bar{E}_{MM} is the average molecular mechanics energy, the term $3RT$ corresponds to the enthalpy of the six translation and rotational degrees of freedom in the classical limit, $\bar{G}_{\text{PB-elec}}$ is the electrostatic solvation energy obtained from Poisson–Boltzmann calculations,²⁶ $\bar{H}_{\text{vdW solute-solvent}}$ is the solute–solvent van der Waals energy contributing to the nonpolar part of solvation energy,²⁵ and $-T\bar{S}_{\text{MM-GBSA}}^{\text{norm}}$ is the solute entropy as estimated by molecular mechanics normal mode calculations and standard statistical mechanical formulas. Equation 1 lacks the cavitation free energy contribution (G_{cav}) to the nonpolar solvation energy, which is usually included in the MM-PB energy expressions as a simple term proportional to the molecular surface area, because we employed in this work an alternative approach for computing G_{cav} (see below).

The SANDER program was used to compute (no cutoff) the molecular mechanics energy terms (E_{MM}). The electrostatic contributions to the solvation free energy ($G_{\text{PB-elec}}$) were determined using the PBSA program included in the Amber 9 package. In the PB calculations, atomic charges and radii were taken from the AMBER03 representation. The linearized PB equation was solved on a cubic lattice by using an iterative finite-difference method. The cubic lattice had a grid spacing of 0.50 Å, and the points at the boundary of the grid were set to the sum of Debye–Hückel potentials. The $\bar{H}_{\text{vdW solute-solvent}}$ terms were determined for a water shell of 12 Å thickness around the solute molecules with no cutoff using SANDER.

Solute entropic contributions were estimated using the NAB package.²⁷ Prior to the normal mode calculations, the geometries of the systems described by their AMBER03 representations were minimized until the root-mean-squared deviation of the elements in the gradient vector was less than 10^{−5} kcal/(mol Å). These minimizations and the subsequent normal mode calculations²⁸ were carried out using the HCT GB model for representing solvent environment. In this work, free energies were computed for a standard state of 0.001 M that is similar to the POG10 concentration employed in DSC experiments⁸ (1.8 mM.). Hence, the translational entropy is 7.4 cal mol^{−1} K^{−1} larger than the entropy value obtained for the standard state of

an ideal gas, owing to the change in concentration from 0.045 M (ideal gas) to 0.001 M (solution).

Cavitation Free Energy. In the last few years, different theoretical analyses as well as free energy perturbation (FEP) or thermodynamic integration (TI) calculations,²⁹ have shown that the free energy required to create a cavity within solvent in order to accommodate the solute (i.e., G_{cav}) is proportional to the solvent-excluded volume (V) for small solutes, whereas for larger sized solutes G_{cav} is proportional to the surface area of the solute (A), the turning point from V - to A -dependence appearing in the range of effective spherical radii of ~ 5 – 10 Å. Since this is precisely the size range corresponding to the POG10 and T3–785 monomers, it seems reasonable that the computation of G_{cav} should take into account both V - and A -effects.

On the basis of the work by Höfner and Zerbetto, who have performed FEP and TI calculations of the G_{cav} energies that are required to create spherical cavities in water with different effective radii (R),^{30,31} G_{cav} fits well to a quadratic function:

$$G_{\text{cav}}^{\text{rPA}}(R) = k_0 + k_1 R + k_2 R^2 \quad (2)$$

where R is the cavity radius in Å, and k_0 , k_1 , and k_2 are fitting parameters ($k_0 = 0.427$ kcal/mol, $k_1 = -1.594$ kcal/(mol Å), and $k_2 = 1.183$ kcal/(mol Å²) for water at 300 K).³² This expression is named as the “revised Pierotti approach” (rPA), given that it closely resembles the formula of G_{cav} derived analytically by Pierotti within the context of the scale particle theory of liquids.³³ Most interestingly, Höfner and Zerbetto have also shown that essentially the same G_{cav} values are obtained when several adjoining spherical cavities are replaced by a single cavity of the same volume. This means that G_{cav} is nearly additive in terms of molecular volume, thus justifying the use of a single spherical “representative” cavity to describe molecules of varying shape, but the same volume.

In principle, the applicability of the rPA formula is limited to the domain of the reference FEP G_{cav} data ($R < 5.0$ Å).³² For larger systems, the relatively large statistical uncertainty of the FEP calculations together with worse fitting parameters can result in absolute errors of tenths of kcal/mol. Therefore, the use of expression 2 for the computation of the G_{cav} energies of the POG10 and T3–785 systems is precluded by the large size of these molecules, which have effective radii of ~ 9 and ~ 13 Å in their monomeric and triple-helical forms, respectively. This problem can be circumvented by computing the change in G_{cav} upon the formation of the triple-helical complex instead of computing the absolute G_{cav} values. To this end, we first computed the change in the 3P \rightarrow TH transition of the solvent-excluded molecular volume, $\Delta V = +86$ and $+135$ Å³ for POG10 and T3–785, respectively. Thus, by taking advantage of the nearly additive character of G_{cav} , the ΔG_{cav} for the 3P \rightarrow TH process could be estimated from the G_{cav} data of a representative spherical cavity having an effective volume equal to ΔV . However, the formation of the triple-helix also implies a significant decrease of the solvent-excluded molecular surface area, $\Delta A = -225$ and -783 Å² for POG10 and T3–785, respectively. In this case, the concomitant decrease in the cavitation free energy could be estimated by computing the G_{cav} of a second representative sphere with surface area equal to ΔA . In consonance with expectations, the opposite trend in ΔV and ΔA suggests that neither ΔV - nor ΔA -dependent G_{cav} values alone can describe the total ΔG_{cav} change. Thus, for the particular case that an asymmetrical change in the molecular

volume ($\Delta V > 0$) and surface ($\Delta A < 0$) takes place, we propose to estimate the total ΔG_{cav} value by means of the following expression:

$$\Delta G_{\text{cav}} = \frac{R_{\Delta V}}{R_{\Delta A}} G_{\text{cav}}^{\text{rPA}}(R_{\Delta V}) - \left(1 - \frac{R_{\Delta V}}{R_{\Delta A}}\right) G_{\text{cav}}^{\text{rPA}}(R_{\Delta A}) \quad (3)$$

where $R_{\Delta V}$ and $R_{\Delta A}$ are two effective radii, which are defined through the changes in the molecular surface area ($R_{\Delta A} = (|\Delta A|/4\pi)^{1/2}$) and volume ($R_{\Delta V} = (3\Delta V/4\pi)^{1/3}$). Note that the $1 - (R_{\Delta V}/R_{\Delta A})$ index quantitatively characterizes the shape of the cavity change, measuring the degree of its deviation from a sphere.³⁴ All the solvent-excluded molecular V and A values were determined using the MSMS program³⁵ with a probe sphere of 1.4 Å radius. The van der Waals radii for the peptide atoms were taken from the united-atom AMBER-84 force-field.

Conformational Entropy. The total configurational entropy of a solute molecule can be estimated from a MD simulation by means of the following approximation:³⁶

$$S_{\text{total}} = \bar{S}_{\text{vib}} + S_{\text{conf}} \quad (4)$$

where \bar{S}_{vib} is the average vibrational entropy as evaluated over a set of representative snapshots by means of normal model calculations, and S_{conf} is the conformational entropy of the whole MD trajectory, which, in turn, can be computed using the Shannon entropy or information entropy:

$$S_{\text{conf}} = -R \sum_{\alpha} p_{\alpha} \ln p_{\alpha} \quad (5)$$

where the index α runs over all possible conformers of the solute molecule and p_{α} is the statistical weight of the α -conformer. Each conformer can be univocally labeled with an array $\{A_i\}_i = 1, N$, where N is the number of torsion angles of the molecule, and A_i is an index that defines the conformational state (e.g., $g+$, $g-$, $anti$) of the i th torsion angle. Due to the huge number of accessible α -conformers ($\sim 3^N$), it is almost impossible to explore *all* the conformational space of biomolecules during typical MD simulations. However, based on the converged probability distributions of the individual torsion angles, it is possible to estimate the conformational entropy of a solute molecule by expanding the global informational entropy of eq 5 in terms of the so-called mutual information functions (MIF),³⁷

$$S(A_1 \dots A_N) = \sum_{i=1}^N S(A_i) - \sum_{i < j} I_2(A_i, A_j) + \sum_{i < j < k} I_3(A_i, A_j, A_k) - \sum_{i < j < k < l} I_4(A_i, A_j, A_k, A_l) + \dots \quad (6)$$

where $S(A_i)$ is the informational entropy of the i th torsion angle. The second-order MIF is computed by combining the informational entropies of single- and two-variable probability distributions of torsion angles:

$$I_2(A_1, A_2) = S(A_1) + S(A_2) - S(A_1, A_2) \quad (7)$$

Similarly, the third-order MIF includes the correlations among three torsion angles:

$$I_3(A_1, A_2, A_3) = S(A_1) + S(A_2) + S(A_3) - S(A_1, A_2) - S(A_1, A_3) - S(A_2, A_3) + S(A_1, A_2, A_3) \quad (8)$$

and the n -order MIF is given by:

$$I_n(A_1, \dots, A_n) = \sum_{k=1}^n (-1)^{k+1} \sum_{i_1 < \dots < i_k} S(A_{i_1}, \dots, A_{i_k}) \quad (9)$$

where the sum $\sum S(A_{i_1}, \dots, A_{i_k})$ runs over all possible combinations $\{i_1, \dots, i_k\} \in \{1, \dots, n\}$.

Results

MD Simulations of the Triple-helices: Global Motions, Helical Structure, and Interchain Contacts. Starting at the corresponding X-ray structures,^{2,18} the POG10 and T3-785 triple-helices were subject to molecular dynamics (MD) simulations (20 ns) at constant P (1 atm) and T (300 K) in explicit solvent using the AMBER package.¹⁹ The MD models were equilibrated with respect to the total root mean squared deviations (rmsd) after ~ 5.0 ns of simulation time. During the last 15 ns of the trajectories, the average rmsd values, ~ 3.0 and ~ 2.8 Å for POG10 and T3-785, respectively, indicate that the THP models deviate moderately from the crystal structures (see Table 1). The three central [(X-Y-Gly)₃] triplets of each THP molecule remain structurally closer to the X-ray structures, as they have rmsd values of 1.37 and 0.66 Å for POG10 and T3-785, respectively. The larger POG10 rmsd values are well-explained by the fact that each polypeptide chain in the starting X-ray model (1CAG) has a central Pro-Hyp-Ala triplet instead of Pro-Hyp-Gly.

To analyze the flexibility of the THP systems, we performed principal component analyses (PCA) of the C α covariance matrices (see Table 2). The first five PCA modes account for $\sim 80\%$ of the structural variance of the backbone C α atoms. Half of this structural variability is described by the two most important PCA modes, which correspond to orthogonal bending motions of the whole THP backbone. The third PCA component, which accounts for only 8% of the internal motions of the C α atoms, corresponds to twisting motions of the triple-helices. Visualization of the projection of the PCA modes onto the MD trajectories shows that the essential dynamics of the two THP molecules is very similar.

By plotting the rmsd values and the PCA eigenvalues of the bending motions (see Supporting Information Figure S2), it is shown that the larger rmsd values the larger amplitude of the bending motions. This means that the structural deviations of the THPs in aqueous solution with respect to the solid state structures are largely due to the bending motions of the THP backbone atoms which, in turn, arise from small fluctuations ($\pm 10^\circ$) of the consecutive Ψ and Φ torsion angles along each peptide chain. In fact, the backbone conformation remains perfectly stable in the aqueous solution, excepting those of the terminal end residues (see Supporting Information Figure S3).

The basic helical parameters that are commonly used to describe the structure of collagen molecules are the unit twist angle (θ) and the unit height (h). We first computed the θ and h parameters for the positions included in the central region of the POG10 model following the prescriptions described in the Supporting Information (see Figure S4). After averaging the data for each MD snapshot and then all along the trajectory, the resulting values are $\theta = 55 \pm 10^\circ$ and $h = 8.8 \pm 0.10$ Å. For the T3-785 model, the corresponding values are $\theta = 44$

TABLE 1: RMSD (Å) over All Heavy Atoms and over the Backbone^a

model	all		central zone	
	T3-785	POG10	T3-785	POG10
rmsd (backbone)	2.43 \pm 0.42	2.66 \pm 0.39	0.66 \pm 0.11	1.37 \pm 0.14
rmsd (all heavy)	2.76 \pm 0.44	2.99 \pm 0.43	1.23 \pm 0.16	1.60 \pm 0.18

^a The central zone comprises three (X-Y-Gly)₃ units.

TABLE 2: Percentage of Variance Accounted for and Physical Description of the Five Most Important Principal Components Obtained after Diagonalization of the C α Covariance Matrix for the THP in the POG10 and T3-785 Trajectories.

mode	% variance accounted for		PCA mode description
	POG10	T3-785	
1	32.2	31.9	global bending
2	28.0	25.3	global bending
3	8.8	8.8	helical twist
4	6.9	7.1	double bending
5	6.3	5.9	double bending

$\pm 10^\circ$ and $h = 9.0 \pm 0.11$ Å. These values are similar to those obtained in the X-ray structures: $\theta = 51.8^\circ$ and $h = 8.73$ Å for the prototypical structure (1VH7), and $\theta = 48$ and $h = 8.9$ for the T3-785 structure. Note that the 1VH7 structure was refined as an infinite helix model with a 7/2 symmetry. In Figure 1 we compare the θ values for the different positions along the triple-helix. For the POG10 model, the helical structure of the whole THP molecule is almost perfectly regular, in accordance with its repeating sequence. The corresponding θ values of the X-ray structure of the Gly \rightarrow Ala mutant matches closely the POG10 data in the prototypical [(Pro-Hyp-Gly)₃] units, but it shows a marked difference in the central region containing the Ala residues at the Gly positions. On the other hand, the triple-helical structure of the T3-785 model is modulated by its peptide sequence so that the THP regions poor in imidic acids are partially unwound ($\theta = \sim 35^\circ$) with respect to the end regions that contain prototypical [(Pro-Hyp-Gly)₃] units. The sequence dependence of the twist angle that has been determined experimentally is satisfactorily reproduced by the MD simulations in aqueous solution.

Other structural features characteristic of the collagen triple-helix, such as the ring-puckering of imidic residues and the interchain H-bonds, are also well-described by the MD simulations (see Supporting Information Figure S5 and Tables S1-S3). For example, the interchain Gly-NH \cdots O=C-X H-bond maintains a relatively short distance between heavy atoms (~ 3.0 Å) and has a large % of abundance (90-100%) during both the POG10 and the T3-785 MD simulations. Besides the Gly-NH \cdots O=C-X interactions, THP molecules in which the X position is occupied by amino acids different from Hyp or Pro, can present further H-bond interactions among the peptide chains mediated by water molecules. In the case of the T3-785 peptide, its central region shows nine Gly-C=O \cdots (H₂O) \cdots HN-X interactions as determined crystallographically. During the T3-785 simulation, the interchain water bridges are fluxional in the sense that other water molecules can diffuse in and replace the existing water molecules in the bridge. The abundance of one-water bridges is generally high ($\sim 70\%$), but their average lifetime is limited to 8-10 ps. These results are in consonance with NMR data, revealing that the first solvent shell around the THP molecules, which include the water bridges and all the water molecules in the first hydration sphere around the protein

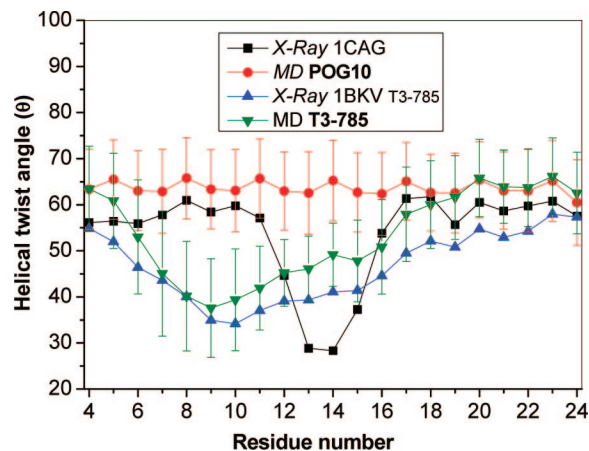


Figure 1. Helical twist angle along the POG10 and T3-785 triple-helices. Vertical bars represent the standard deviation of the MD values.

atoms, is kinetically labile with upper limits for water molecule residence times in the nanosecond to subnanosecond range.³⁸

MD Simulations of the Monomeric Forms. The problem of finding out the structure of the monomeric forms of POG10 and T3-785 can be seen as a polypeptide folding problem that, in turn, can be solved computationally for small peptides (6–15 residues) through the combination of accurate force fields and extensive MD samplings.³⁹ A similar computational approach is applied here because the POG10 and T3-785 chains, which are 30 residues-long, are relatively rigid due to their high content of imidic acid residues. We also assume that the free chains keep all their peptide bonds in the most stable *trans* conformation.

A 50 ns MD of a single POG10 chain, started at an extended conformation, rapidly evolved toward a compact structure, with its radius of gyration decreasing abruptly from 16 to 9–8 Å at 12 ns. To further assess the stability of this compact structure, we carried out a conformational search using the LMOD algorithm.²¹ The set of minimum energy structures predicted by LMOD turned out to be very close both in energy (2–3 kcal/mol) and in structure (rmsd values <1.0 Å) to the 50 ns structure (see Supporting Information Figure S6), suggesting thus that the POG10 folding was robust. This was further confirmed by a second 50 ns simulation started at the most stable LMOD structure, from which snapshots were extracted for analyses and free energy calculations. During this simulation, the POG10 folding is characterized by three turn segments (Pro₇-Pro₁₀; Pro₁₃-Pro₁₆ and Hyp₂₀-Hyp₂₃) stabilized by 9 intramolecular C=O⋯HN interactions between backbone groups with occupation higher than 80%.

Following the same computational protocol that was used for the POG10 peptide, we observed that the T3-785 peptide folds after 10 ns into a compact form, but its radius of gyration fluctuates widely. The stability of the 50 ns snapshot was examined by conformational search calculations. The resulting LMOD structures exhibited significant structural and energetic variability, thus suggesting that T3-785 could have an important conformational flexibility. Hence, we run a second MD simulation from the most stable LMOD structure that extended up to 250 ns. All along this second simulation, the backbone chain of T3-785 experiences frequent conformational changes, as shown by clustering analyses (see Figure 2). Only the last 150 ns of this simulation were considered for structural and energetic analyses. Secondary structural analyses on selected MD snapshots lead preferentially to a coil assignment for all the T3-785 residues, although two short segments Pro₄-Pro₇ and Gly₁₅-Gly₁₈

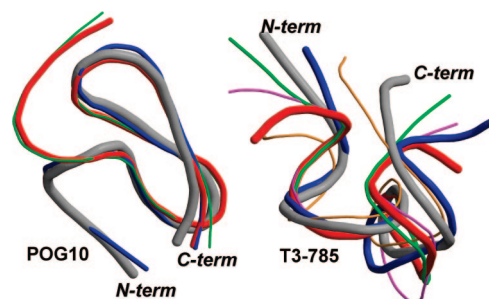


Figure 2. Superposition of the most populated representative structures derived from clustering analyses for the POG10 and T3-785 monomers. Thickness of the models corresponds to the number of snapshots represented by each model.

TABLE 3: Average Values^a of the Free Energy Components for the Transition from the Monomeric to the Triple-helix State at 300 K

	POG10		T3-785	
	mean	standard error	mean	standard error
$\Delta \bar{E}_{\text{MM}}$	81.2	1.6	74.2	1.4
$\Delta \bar{G}_{\text{PB-elec}}$	-112.3	1.2	-112.0	1.0
$\Delta \bar{H}_{\text{vdW solute-solvent}}$	16.2	0.6	28.6	0.6
$\Delta \bar{G}_{\text{cav}}$	-0.7	0.1	-11.3	0.2
$-T\Delta \bar{S}_{\text{MM-PB}}^{\text{GBSA}}$	8.1	0.1	14.6	0.2
$\Delta \bar{G}_{\text{MM-PB}}$	-6.6	1.2	-6.1	1.1
$-T\Delta \bar{S}_{\text{conf}}$	0.4		4.6	
$\Delta \bar{G}_{\text{total}}$	-6.2	1.2	-1.5	1.1

^a kcal per mol of peptide.

adopt a turn conformation with about 25 and 50% occupation, respectively.

MM-PB Free Energy Calculations. Molecular mechanics (MM), electrostatic Poisson-Boltzmann (PB), and normal mode MM calculations²⁵ were carried out on MD snapshots of the POG10 and T3-785 peptides in their monomeric (P) and triple-helix (TH) states. The nonpolar contribution to the solvation energy was obtained by combining the enthalpy of the van der Waals interaction between the solute and a 12 Å shell of water molecules, with the cavitation free energy (G_{cav}) of the solute as estimated by molecular volume and surface area calculations as described above. From these calculations, we obtained average $\Delta G_{\text{MM-PB}}$ energies for the transition from monomer to triple-helix (i.e., P → 1/3 TH). We see in Table 3 that all the energetic terms contributing to $\Delta G_{\text{MM-PB}}$ vary significantly upon going from P to TH. Interestingly, the driving force for the formation of the triple-helix is mainly provided by the electrostatic solvation energy, thus showing that the loss of intramolecular interactions in the P state is not compensated by interchain interactions.

From the DSC curves of the POG10 peptide,⁸ Nishi et al. have derived thermodynamic parameters for the P → 1/3 TH process, including the difference of C_p before and after the transition. On the basis of their data, we obtain an experimental ΔG value at 300 K of -6.4 kcal/mol. The corresponding $\Delta G_{\text{MM-PB}}$ value amounts to -6.6 kcal/mol with a statistical uncertainty of 1.2 kcal/mol (standard error), which is in reasonable agreement with the experimental value. For the T3-785 system, however, its average $\Delta G_{\text{MM-PB}}$ of -6.1 kcal/mol is quite close to that of POG10, which does not explain the much lower thermal stability of the T3-785 triple-helix as compared with POG10.

Conformational Entropy Calculations. Most likely, the similar stability of POG10 and T3-785 in terms of their $\Delta G_{\text{MM-PB}}$ values is due to the fact that normal mode MM calculations

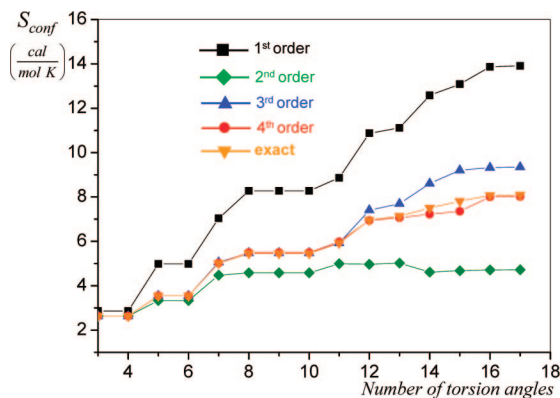


Figure 3. Conformational entropies (cal/mol K) of torsion angle subsets of varying size as computed with expansions of MIFs of 1st–4th-order. All the calculations were done on 15 000 snapshots taken from a trajectory segment of the MD simulation of T3–785 in its monomeric form.

do not include the conformational entropy (S_{conf}) that accounts for the number of significantly occupied conformers of solute molecules. To derive S_{conf} values from our MD simulations, we employed an expansion of the so-called mutual information functions (MIF)^{37,40} that approaches the full-dimensional conformational probability distribution by including systematically N -order correlations among the internal degrees of freedom.

To make the S_{conf} calculations feasible, it was necessary to reduce the dimensionality of the problem as well as to truncate the expansion of the MIFs. Taking into account that the largest conformational changes occurring in the aggregation process correspond to the backbone chains, we computed the S_{conf} values arising from the conformational freedom of the Φ and Ψ torsion angles. In addition, the expansion of MIFs was truncated to fourth-order, as suggested by preliminary test calculations on the isolated T3–785 peptide (see Figure 3). For such calculations, torsion angles were chosen starting at the N -terminal end, and the corresponding conformational probability distributions were computed for a 15 ns trajectory segment. The fourth-order expansion of MIFs is quite close to the “exact” full conformational entropy for a number of torsion angles ranging from 3 to 17. The first-order expansion of MIFs, which assumes that all torsion angles are independent variables, largely overestimates the exact conformational entropy. Inclusion of pair correlation always results in entropy values lower than the exact ones, but, in general, the computed entropy does not converge monotonically toward the exact value when including higher order MIF terms.³⁷

Figure 4 shows the convergence plots of S_{conf} for all the systems examined in this work. By using a 15 ns MD trajectory segment, the fourth-order MIFs expansion approaching the S_{conf} data for the backbone torsion angles results in small and reasonably converged entropy values for the two POG10 systems and the T3–785 triple-helix state. The MIF expansion for the triple-helix states results in $-TS_{\text{conf}}$ values of about -1.0 kcal/mol. Similarly, the $-TS_{\text{conf}}$ term of the POG10 monomer is not large, -0.6 kcal/mol. Moreover, we see in Figure 2 that the first-order S_{conf} entropy computed by assuming independent motions of the Φ and Ψ torsion angles is very similar to the fourth-order correlated data. This is not entirely unexpected given that the backbone conformation is basically frozen in both the POG10 and T3–785 triple-helices, as well as in the POG10 monomer. The backbone chain of the T3–785 monomer exhibits a complex dynamic behavior, as suggested by the clustering analyses. For this system, the first- and fourth-order S_{conf} values

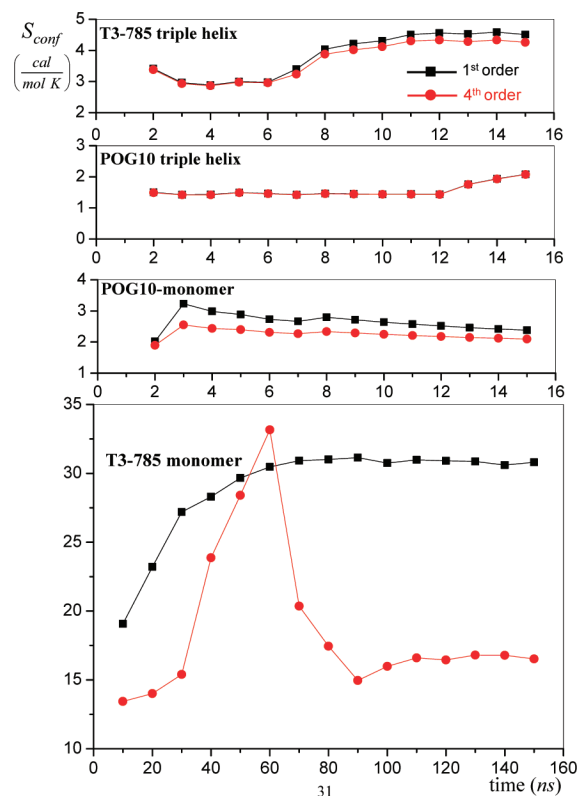


Figure 4. Convergence plots of the conformational entropy (cal/mol K) for POG10 and T3–785 in their monomeric and triple-helical states. Entropies derived from MIF expansions of first-order (black boxes and lines) and fourth-order (red circles and lines) are plotted for each system.

differ significantly all along the trajectory, thus showing the importance of correlation between motions along the torsion degrees of freedom. The converged S_{conf} value for the T3–785 peptide gives a free energy term $-TS_{\text{conf}}$ of -5.0 kcal/mol, which is about 10 times larger than that of POG10.

Owing to the relatively large flexibility of the isolated T3–785 peptide, a broad sampling of the conformational space (150 ns) was required to obtain converged S_{conf} values. During the first half of the analyzed trajectory, the T3–785 peptide explores constantly new conformations that induce a significant increase in entropy resulting in a broad peak at 60 ns. Afterward, the T3–785 peptide remains basically in the same set of conformations for the rest of the trajectory, which reduces its entropy until it reaches a plateau. By means of clustering analyses, we confirmed that, during the second part of the trajectory, the conformational space explored by the T3–785 peptide is quite similar to that sampled during the first 75 ns (see Supporting Information Figure S7). These analyses suggest that the backbone conformational variability of the T3–785 peptide is reasonably sampled and that the corresponding S_{conf} plot in Figure 4 is converged. We also note that peaks in convergence plots of S_{conf} due to sampling issues have also been observed in previous calculations of absolute entropies for polypeptides using the Schlitter method.⁴¹

Overall, the conformational entropy penalty to the P \rightarrow 1/3 TH process is 0.4 and 4.6 kcal/mol for POG10 and T3–785, respectively. By combining these values with the MM-PB data, the total ΔG change has now average values of -6.2 and -1.5 kcal/mol. Therefore, we conclude that the lower thermal stability of the T3–785 triple-helix is basically determined by the larger conformational entropy of its monomer state.

Discussion

The MM representations of the THPs reproduce quite well both the global and local structural features that are characteristic of the collagen triple-helix. Moreover, the MD simulations complement the X-ray and NMR data by providing new information about the flexibility of these molecules in solution. The simulations can also predict the molecular properties of the isolated peptides in aqueous solution, which are usually considered as random-coil states. For the prototypical POG10 peptide, however, its high content of imidic acids leads to a quite stable backbone conformation. Clustering analyses clearly suggest that the isolated POG10 peptide is better described as a folded structure instead of a random-coil. The T3–785 peptide, which is poor in imidic acids, adopts a partially folded structure, although it is clear that its backbone chain exhibits a large degree of conformational variability. In fact, a very long simulation (up to 250 ns) was required for proper equilibration and sampling.

From the MD trajectories we obtained meaningful and reasonably accurate free energy changes: -6.2 and -1.5 kcal/mol, for the formation of the POG10 and T3–785 triple-helices, respectively. Of particular importance in the computational protocol are (a) the computation of the nonpolar solvation energy by combining the explicit solvent representation with an estimation of the relative change in the cavitation free energy of the solute, and (b) the partition of the configurational entropy of the solute into vibrational and conformational contributions that were estimated by normal mode calculations and the MIF approach, respectively. In addition, the free energy calculations can help us to better understand the stability of the THPs. Thus, to date the most commonly accepted view emphasizes the role played by the steric restrictions imposed by imidic acid rings in favoring interchain interactions.⁶ This picture can now be complemented with our results for POG10 and T3–785 showing that the balance of solute–solvent interactions largely determines the global thermodynamic stability of the triple-helical state. Most interestingly, the calculations also suggest that, besides their well-known structural role, the presence of the Pro and Hyp rings is crucial for reducing the entropic penalty due to the loss of conformational freedom during the aggregation process. Replacement of Pro/Hyp by other nonimidic residues would increase the entropy cost, thereby destabilizing the triple-helix. Hence, knowledge of the structure and conformational entropy of the peptide monomers seems to be required to understand the relationship between peptide sequence and thermal stability of THPs.

In retrospect, the fact that the loss of conformational entropy upon triple-helix formation penalizes more severely the more flexible T3–785 molecules than the more rigid POG10 ones is far from surprising, although it has been largely unnoticed. The actual problem was more about how to compute in a balanced manner all the free energy components controlling the absolute and relative stabilities of the THP systems. Nevertheless, our results also remark the convenience of computing the configurational entropy of complex molecular systems as previously noticed by van Gunsteren and co-workers in their analyses of the folding behavior of polipeptides,⁴¹ or by Gilson and co-workers in their studies on the entropic effects of protein–ligand binding.^{40,42} In these and other studies, the configurational entropy of biomolecules has been computed by applying different methodologies on the output provided by extensive MD- or Monte Carlo-based simulations. Similarly, we show that the popular MM-PB computational protocol, which

approximates the configurational entropy by the average vibrational entropy, can be corrected a posteriori with an estimation of the conformational entropy based on information theory. This approach could be applied straightforwardly to compute free energies of other processes involving biomolecular association and/or polypeptide folding.

Acknowledgment. This research was supported by the following grants: FICYT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. N.D. also thanks MEC for her Ramon y Cajal contract.

Supporting Information Available: Figures S1–S7 and Tables S1–S3 (12 pages) as noted within the text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References and Notes

- Rich, A.; Crick, F. H. C. *J. Mol. Biol.* **1961**, *3*, 483.
- Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75.
- Wess, T. J. *Adv. Protein Chem.* **2005**, *70*, 341.
- Orgen, J. P. R. O.; Irving, T. C.; Miller, A.; Wess, T. J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9001–9005.
- Leikina, E.; Merts, M. V.; Kuznetsova, N.; Leikin, S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1314–1318.
- Collagen: Primer in Structure, Processing and Assembly*; Brinckmann, J.; Notbohm, H.; Müller, P. K., Eds.; Springer-Verlag: Berlin-Heidelberg, 2005; Vol. 247.
- Brodsky, B.; Persikov, A. V. *Adv. Protein Chem.* **2005**, *70*, 301.
- Nishi, Y.; Uchiyama, S.; Doi, M.; Nishiuchi, Y.; Nakazawa, T.; Ohkubo, T.; Kobayashi, Y. *Biochemistry* **2005**, *44*, 6034.
- Kotch, F. W.; Raines, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3028.
- Cejas, M. A.; Kinney, W. A.; Chen, C.; Vintert, J. G.; Almond, H. R. J.; Bals, K. M.; Maryanoff, C. A.; Schmidt, U.; Breslav, M.; Mahan, A.; Lacy, E.; Maryanoff, B. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 8513.
- Li, M.-H.; Fan, P.; Brodsky, B.; Baum, J. *Biochemistry* **1994**, *32*, 7377.
- Klein, T. E.; Huang, C. C. *Biopolymers* **1998**, *49*, 167.
- Mooney, S. D.; Huang, C. C.; Kollman, P. A.; Klein, T. E. *Biopolymers* **2001**, *58*, 347.
- Mooney, S. D.; Kollman, P. A.; Klein, T. E. *Biopolymers* **2002**, *64*, 63.
- Stultz, C. M. *J. Mol. Biol.* **2002**, *319*, 997.
- Stultz, C. M. *Protein Sci.* **2006**, *15*, 2166.
- Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *14*, 1999.
- Kramer, R. Z.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *Nat. Struct. Biol.* **1999**, *6*, 454.
- Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Matthews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.
- Sherer, E. C.; Harris, S. A.; Soliva, R.; Orozco, M.; Laughton, C. A. *J. Am. Chem. Soc.* **1999**, *121*, 5981.
- Kolossváry, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011.
- Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401.
- Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graphics Model.* **2004**, *22*, 377.
- Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889.
- Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238.
- Sharp, K.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *19*, 301.
- Modeling Unusual Nucleic Acid Structures*; Macke, T.; Case, D. A., Eds.; American Chemical Society: Washington, DC, 1998; pp 379.
- Brown, R. A.; Case, D. A. *J. Comput. Chem.* **2006**, *27*, 1662–1675.
- Chandler, D. *Nature* **2005**, *437*, 640.
- Höfinger, S.; Zerbetto, F. *Chem. Soc. Rev.* **2005**, *34*, 1012.

- (31) Höfner, S.; Zerbetto, F. *Chem.—Eur. J.* **2003**, 9, 566.
(32) Mahajan, R.; Kranzlmüller, D.; Volkert, J.; Hansmann, U. H. E.; Höfner, S. *Phys. Chem. Chem. Phys.* **2006**, 8, 5515.
(33) Pierotti, R. A. *Chem. Rev.* **1976**, 76, 717.
(34) Grigoriev, F. V.; Basilevsky, M. V.; Gabin, S. N.; Romanov, A. N.; Sulimov, V. B. *J. Phys. Chem. B* **2007**, 111, 13748.
(35) Sanner, M. F.; Olson, A. J.; Spohner, J.-C. *Biopolymers* **1996**, 38, 305.
(36) Karplus, M.; Ichiye, T.; Pettit, B. M. *Biophys. J.* **1987**, 52, 1083.
(37) Matsuda, H. *Phys. Rev. E* **2000**, 62, 3098.
(38) Melacini, G.; Bonvin, A. M. J. J.; Goodman, M.; Boelens, R.; Kaptein, R. *J. Mol. Biol.* **2000**, 300, 1041.
(39) Daura, X. *Theor. Chem. Acta* **2006**, 116, 297.
(40) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, 127, 024107.
(41) Schäfer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Proteins Struct. Funct. Genet.* **2001**, 56, 43.
(42) Chang, C. A.; Chen, C.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, 104, 1534.

JP8074699

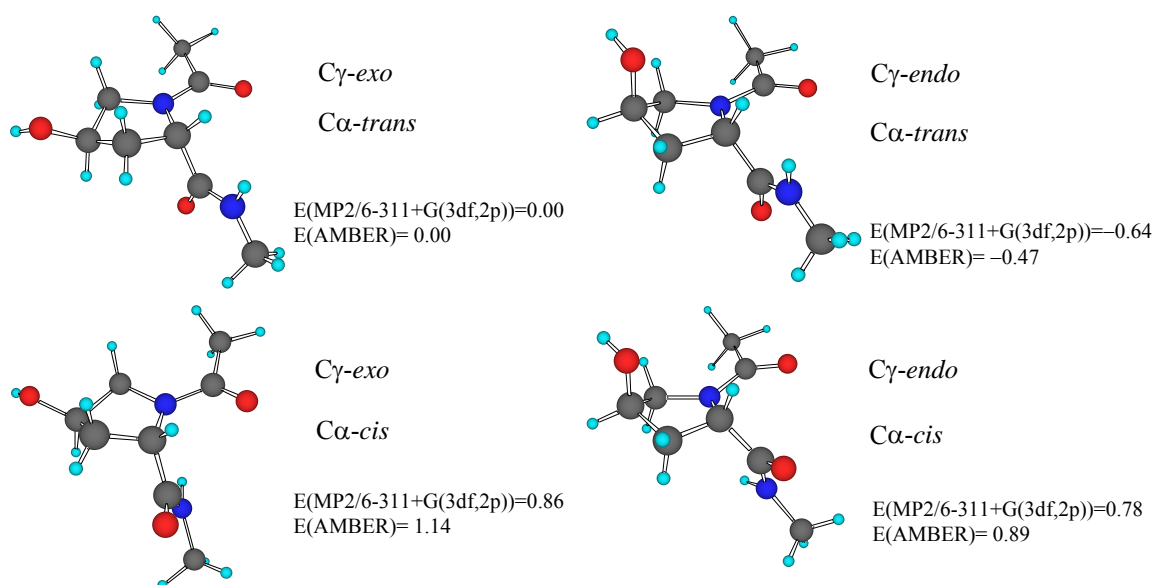
Entropic Control of the Relative Stability of Triple Helical Collagen Peptide Models

Ernesto Suárez, Natalia Díaz and Dimas Suárez*

e-mail: dimas@uniovi.es

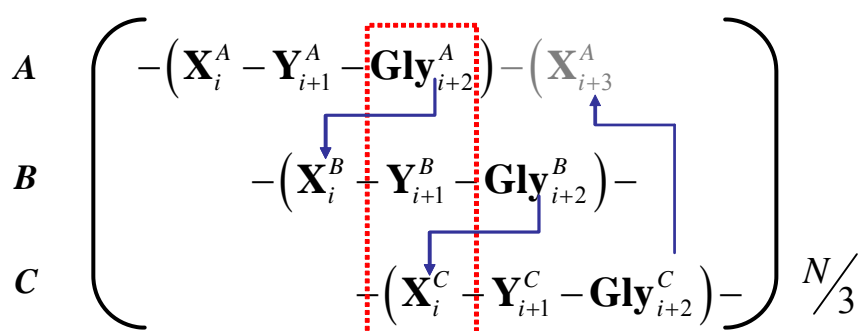
Supporting Information

Figure S1. Views of the Hyp conformers that were considered during parameter derivation. Relative conformational energies in the gas-phase are given in kcal/mol. The Ace-Hyp-Nme conformers were optimized at the HF/6-31G** level using a continuum solvent model. The atomic charges were derived by means of the RESP method¹ using the B3LYP/cc-pVTZ electrostatic potential of the four Hyp conformers. Most of the bond, angle, dihedral and Lennard-Jones parameters of Hyp were available from the AMBER03 database. To better reproduce the correct pucker preference of the five-membered Hyp ring, a specific torsion for the N-C δ -C γ -O atoms was also added. The Hyp parameterization was tested by minimizing in vacuum the geometry of the four conformers.



Labeling of the Peptide Residues in the Triple Helix

The 1CAG and 1BKV X-ray structures contain three peptide chains (labeled as *A*, *B* and *C*) with $N=30$ amino acids per chain (a total of $3N$ residues). The triple helix in these structures is conveniently described as a series of 10 $[(X-Y-Gly)_3]$ units (*i.e.*, $N/3$ units) comprising each one three triplets of residues from the *A*, *B* and *C* chains (see Scheme S1). The fact that the three chains are staggered is seen in that the $C\alpha$ atoms of the Gly_{i+2}^A , Y_{i+1}^B and X_i^C residues lie approximately in a perpendicular plane to the triple helical axis (see the red box in Scheme S1; inter-chain hydrogen bonds are indicated in blue).



Scheme S1

Figure S2. Correlation plot between the norm of the sum of the two first PCA eigenvectors ($C\alpha$ atoms) and the RMSD values for the **POG10** triple helix.

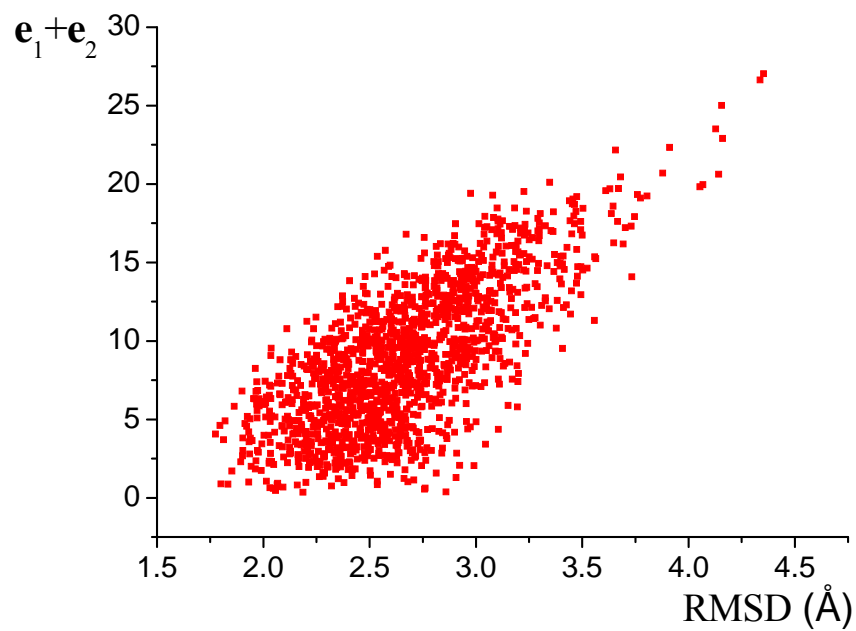
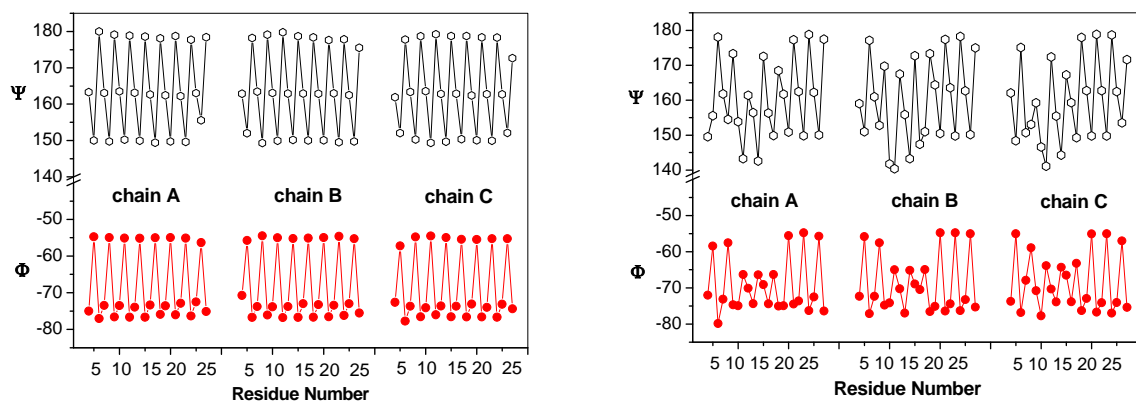


Figure S3. Mean Values of Ψ and Φ angles along the Peptide Chains of the THP models excluding the *N*- and *C*-terminal [(X-Y-Gly)₃] units. For the **POG10** model, the Ψ and Φ values result in a periodic pattern along each of the three peptide chains in consonance with its repeating peptide sequence and its helical symmetry. For **T3-785**, the mean values of the Ψ and Φ angles depend significantly on the exact positioning of the [(X-Y-Gly)₃] triplets in the irregular peptide sequences.



Helical Parameters

The basic helical parameters that are commonly used to describe the structure of collagen molecules are the unit twist angle (θ) and unit height (h). In order to compute these helical parameters from atom coordinates, two algorithms have been proposed in the literature that involve cylindrical² and internal coordinates,³ respectively. In this work, however, the values of θ and h were computed directly from the coordinates of $C\alpha$ atoms. The θ and h values are computed for each residue position along the triple helix. Thus, the twist angle for the i -residue is derived from geometrical data of the three peptide chains.

The helical twist angle (θ) is first computed for the i -residue in chain A using coordinates of $C\alpha$ atoms in the B and C chains as well. The same procedure is applied to compute the θ values for the corresponding residues in the B and C chains, and then the θ values are averaged over the three chains. In our approach, the helical twist angle (θ) on the i -residue is referred to the $(i+3)$ residue and is computed as the dihedral angle between the following two planes:

Plane 1 defined by:

1. Center of mass among the $C_\alpha(i)$, $C_\alpha(i-1)$, and $C_\alpha(i-2)$ atoms in the A , B and C chains respectively.
2. Center of mass among the $C_\alpha(i+3)$, $C_\alpha(i+2)$, and $C_\alpha(i+1)$ atoms in the A , B and C chains. respectively.
3. $C_\alpha(i)$ position in the A chain.

Plane2 defined by:

1. As in *Plane 1*
2. As in *Plane 1*
3. $C_\alpha(i+3)$ position in the A chain.

The unit height (h) is defined as the distance between the points 1 and 2 in *Plane 1*.

Figure S4. Definition of the helical parameters computed in this work. See Scheme S1 for Residue Labeling details.

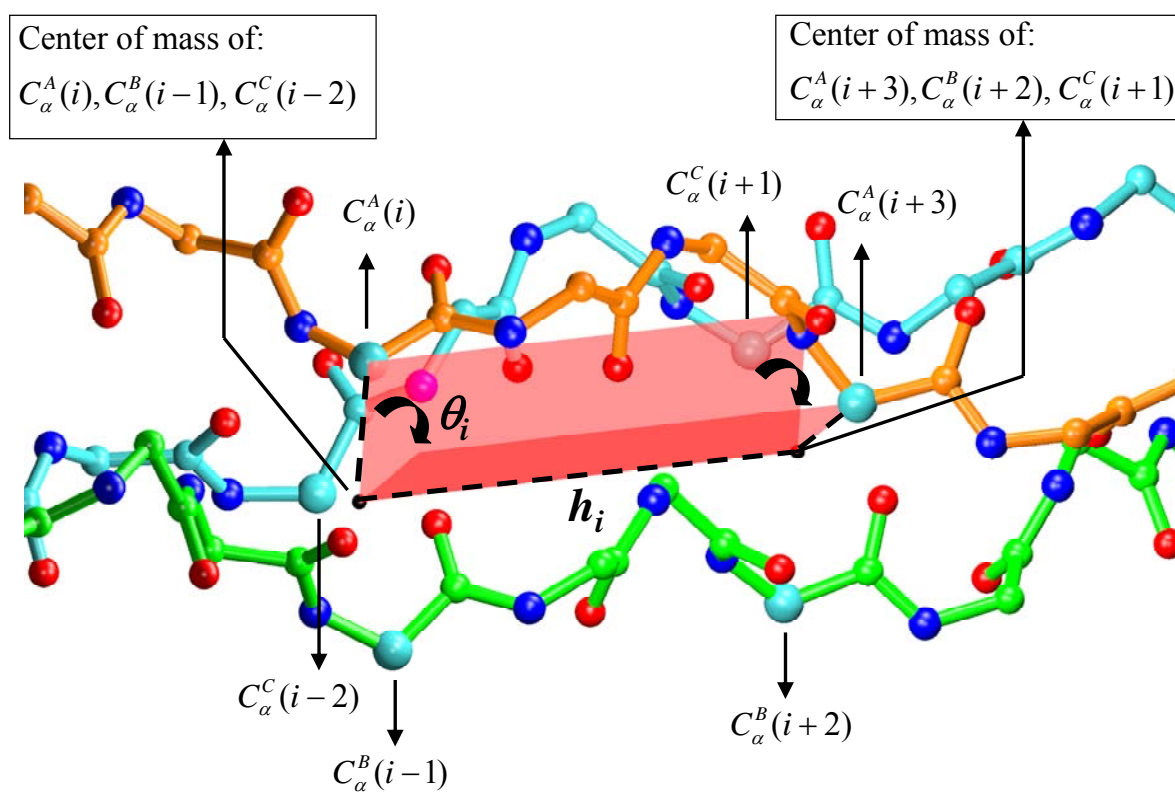


Figure S5. Frequency histograms for the Pro and Hyp ring puckering using the Cremer and Pople definition.⁴ During the simulation of the **POG10** triple helix, the proline ring puckering of the Pro residues located at the X position has a clear C γ -endo preference (~81%). Reciprocally, the large majority of the Hyp residues at the Y position adopt the C γ -exo conformation (97%). The abundance of the Pro-C γ -endo and Hyp-C γ -exo conformers is nearly coincidental in the **POG10** and **T3-785** simulations. This conformation behavior is in agreement with X-Ray⁵ and NMR experiments.⁶

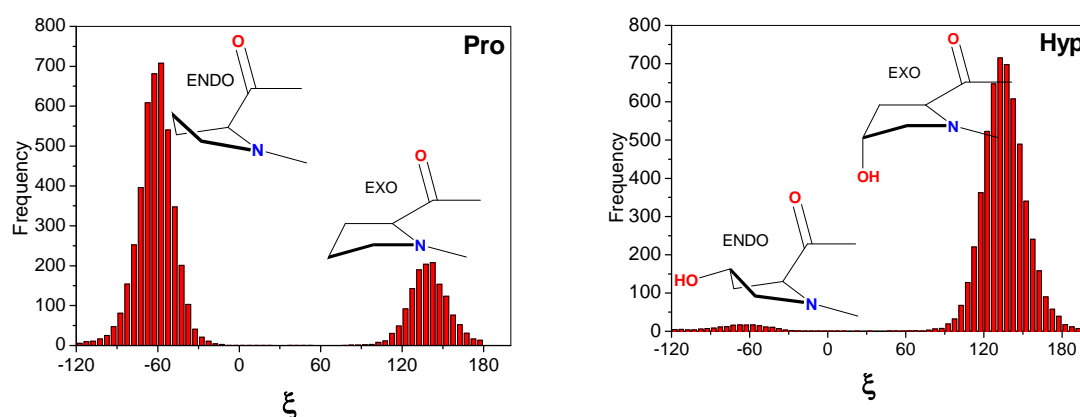


Table S1. Average Distances Between Heavy Atoms (Å) and % of Occurrence Data for Important H-Bond Interactions Stabilizing the Triple Helix.

H-Bonds	T3-785		PROT	
	%	distance	%	distance
Gly6A-NH.....OC-X4B	94.56	3.08±0.18	96.06	3.07±0.17
Gly9A-NH.....OC-X7B	92.32	3.10±0.18	96.70	3.07±0.17
Gly12A-NH.....OC-X10B	98.82	2.95±0.15	96.75	3.07±0.17
Gly15A-NH.....OC-X13B	99.38	2.97±0.15	97.03	3.07±0.17
Gly18A-NH.....OC-X16B	99.23	2.93±0.15	96.53	3.07±0.17
Gly21A-NH.....OC-X19B	97.55	3.06±0.17	96.77	3.07±0.17
Gly24A-NH.....OC-X22B	96.47	3.08±0.17	96.04	3.07±0.17
Gly27A-NH.....OC-X25B	94.96	3.07±0.17	97.18	3.04±0.17
Gly6B-NH.....OC-X4C	95.36	3.08±0.17	96.43	3.07±0.17
Gly9B-NH.....OC-X7C	97.83	3.01±0.17	96.69	3.07±0.17
Gly12B-NH.....OC-X10C	98.91	2.97±0.15	97.03	3.07±0.17
Gly15B-NH.....OC-X13C	99.49	2.97±0.15	96.71	3.07±0.17
Gly18B-NH.....OC-X16C	98.57	2.98±0.16	96.91	3.07±0.17
Gly21B-NH.....OC-X19C	97.05	3.06±0.17	96.52	3.07±0.17
Gly24B-NH.....OC-X22C	97.19	3.07±0.17	96.57	3.07±0.17
Gly27B-NH.....OC-X25C	96.37	3.04±0.17	96.76	2.99±0.17
Gly6C-NH.....OC-X7A	97.39	3.04±0.17	95.73	3.08±0.17
Gly9C-NH.....OC-X10A	99.54	2.90±0.14	96.67	3.07±0.17
Gly12C-NH.....OC-X13A	97.22	3.01±0.16	96.77	3.07±0.17
Gly15C-NH.....OC-X16A	99.37	2.97±0.15	96.96	3.07±0.17
Gly18C-NH.....OC-X19A	90.18	3.13±0.18	97.04	3.07±0.17
Gly21C-NH.....OC-X22A	96.58	3.07±0.17	97.00	3.07±0.17
Gly24C-NH.....OC-X25A	95.77	3.07±0.17	96.37	3.07±0.17
Gly27C-NH.....OC-X28A	87.29	3.03±0.18	95.12	3.07±0.17

Table S2. Average Distances Between Heavy Atoms (Å) and % of Occurrence Data for Important H-Bond Interactions Stabilizing the Side Chain of the Arg residues. The presence of bulky and/or polar amino acid side chains (Ile, Arg) is a characteristic feature of the **T3-785** model. In solution, the Arg side chains establish additional inter-chain H-bond contacts with the backbone carbonyl groups of other peptide chains (e.g., $\text{Arg}_{14}^A - \text{N}\epsilon\text{H}\cdots\text{O}=\text{C} - \text{Arg}_{14}^B$) that are quite stable all along the MD simulations.

H-Bonds	MD		X-Ray
	%	dist(Å)	dist(Å)
Arg14A-N ϵ -H...OC-Arg14B	43.79	2.93±0.16	3.02
Arg14A-N η 1-H...OC-Arg14B	20.29	2.94±0.18	-
Arg14A-N η 2-H...OC-Arg14B	20.39	3.21±0.20	-
Arg14B-N ϵ -H...OC-Arg14C	54.21	2.91±0.16	-
Arg14B-N η 1-H...OC-Arg14C	4.35	2.91±0.16	2.85
Arg14B-N η 2-H...OC-Arg14C	25.87	3.21±0.19	-
Arg14C-N ϵ -H...OC-Ala17A	89.95	2.93±0.17	2.75
Arg14C-N η 1-H...OC-Ala17A	39.51	3.21±0.20	-
Arg14C-N η 2-H...OC-Ala17A	-	-	-

Table S3. Percentage of Occurrence Data and Average Life for Important H-Bond Interactions mediated by Water molecules in the **T3-785** triple helix. The following H-bond criteria were used: (a) distance between the heavy atoms less than 3.5Å; (b) angle H-donor-acceptor less than 60 degrees.

Water Mediated H-Bonds	% Abundance	Average life(ps)
Gly _{9A} -CO...H-O...HN-Ile _{10B}	72.3	8.3
Gly _{12A} -CO...H-O...HN-Ala _{13B}	74.1	6.5
Gly _{15A} -CO...H-O...HN-Leu _{16B}	80.1	8.2
Gly _{9B} -CO...H-O...HN-Ile _{10C}	69.4	6.5
Gly _{12B} -CO...H-O...HN-Ala _{13C}	72.8	5.9
Gly _{15B} -CO...H-O...HN-Leu _{16C}	83.6	8.5
Gly _{6C} -CO...H-O...HN-Ile _{10A}	82.4	10.2
Gly _{9C} -CO...H-O...HN-Ala _{13A}	47.4	3.8
Gly _{12C} -CO...H-O...HN-Leu _{16A}	61.8	5.2

Figure S6. Superposition of backbone atoms of the LMOD structures onto the initial structure (in gray). MM-GBSA energies (in kcal/mol) of the minimized structures and RMSD values in parentheses (Å) of the LMOD structures with respect to the initial one are also indicated.

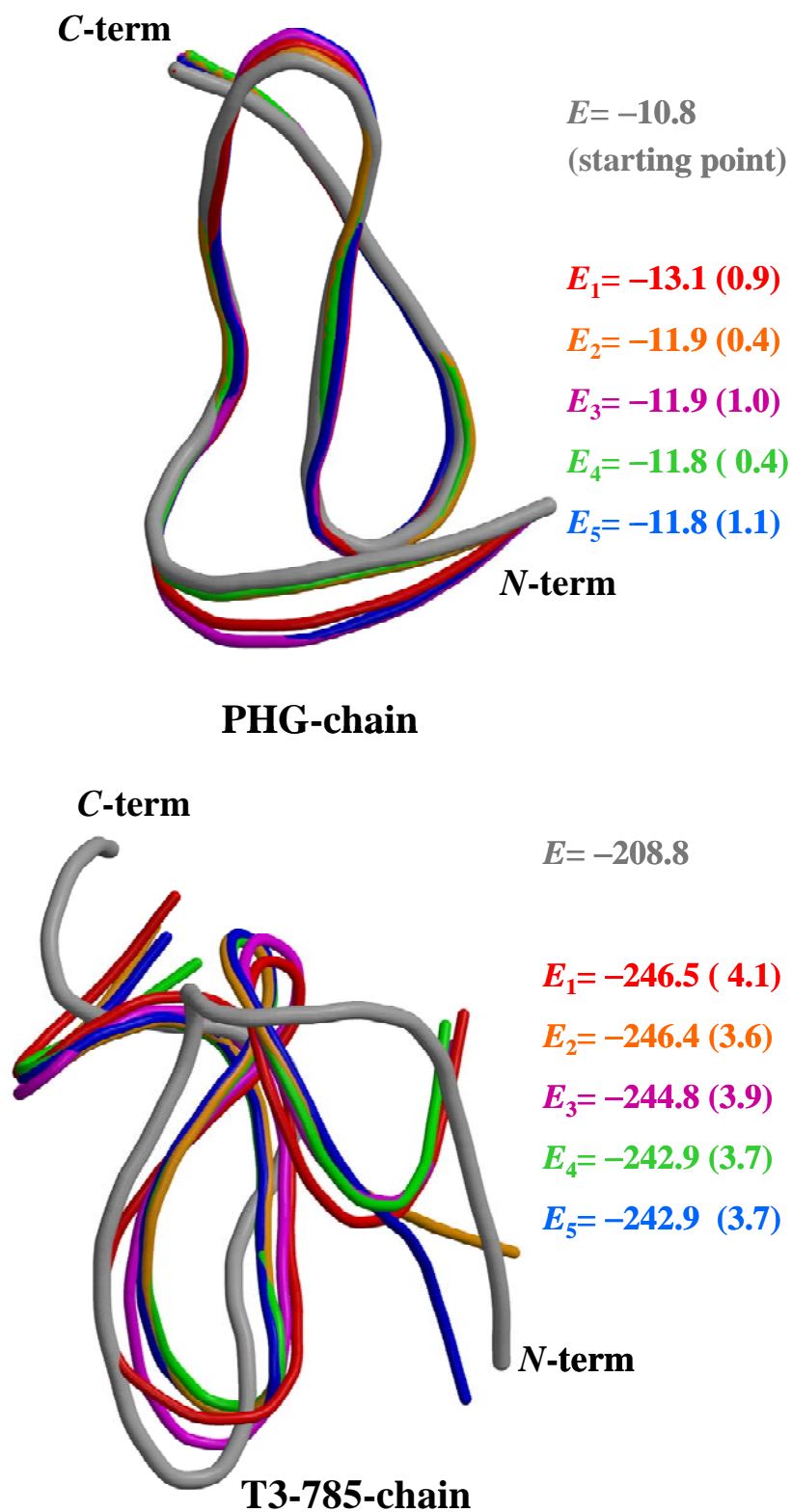
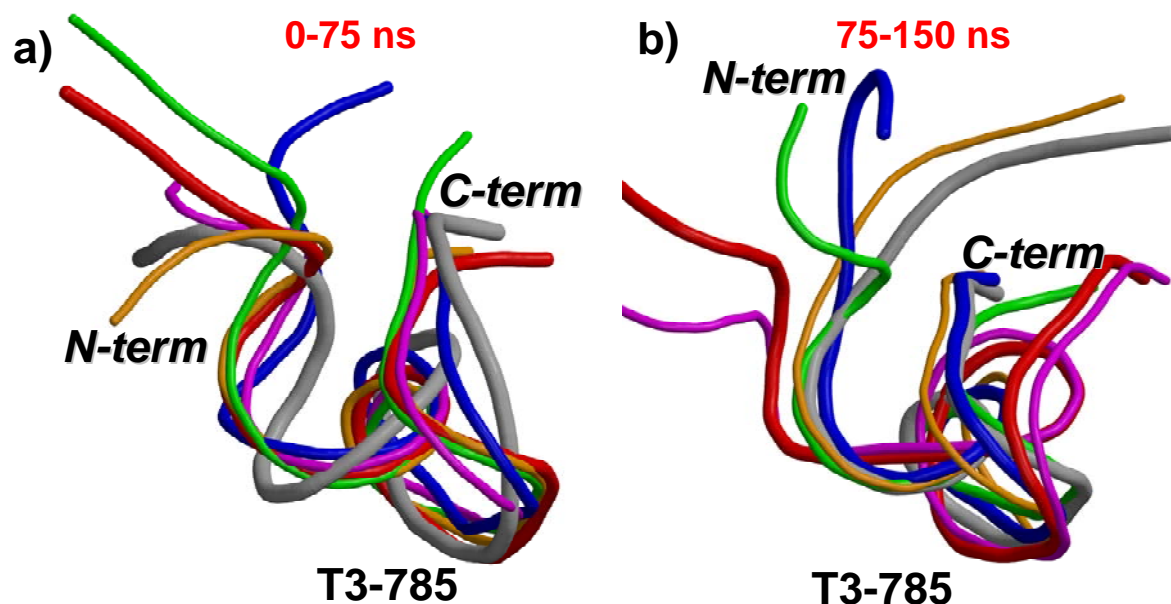


Figure S7. Superposition of the most populated representative structures derived from clustering analyses for the **T3-785** monomer: a) clustering done on the first 75 ns of the production phase; b) clustering done during the second half (75-150 ns). Thickness of the models corresponds to the number of snapshots represented by each model.



References

- (1) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Comput. Chem.* **1993**, *115*, 9620.
- (2) Rainey, J. K.; Goht, C. *Prot. Sci.* **2002**, *11*, 2748–2754.
- (3) Sugeta, H.; Miyazawa, T. *Biopolymers* **1967**, *5*, 673.
- (4) Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354.
- (5) Okuyama, K.; Hongo, C.; Fukushima, R.; Wu, G.; Narita, H.; Noguchi, K.; Tanaka, Y.; Nishino, N. *Biopolymers* **2004**, *76*, 367.
- (6) Li, M.-H.; Fan, P.; Brodsky, B.; Baum, J. *Biochemistry* **1994**, *32*, 7377.

***2.1.1.2 Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix
Metalloproteinase-2: Molecular Dynamics and QM/MM Calculations***

Natalia Díaz, Dimas Suárez and Ernesto Suárez

Proteins 2010, 78: 1-11

Kinetic and binding effects in peptide substrate selectivity of matrix metalloproteinase-2: Molecular dynamics and QM/MM calculations

Natalia Díaz,* Dimas Suárez, and Ernesto Suárez

Departamento de Química Física y Analítica, Universidad de Oviedo, C/ Julián Clavería, 8. 33006 Oviedo, Asturias, Spain

ABSTRACT

Herein, we examine computationally the binding and hydrolysis reaction of the MMP-2 enzyme with two peptide substrates selected by the enzyme from a phage peptide library. Molecular dynamics simulations of the Michaelis complexes (25 ns) allow us to characterize the main enzyme/substrate contacts. Subsequently MM-PBSA calculations using independent trajectories for the complexes and the free substrates provide relative binding energies in good agreement with the experimental K_M results. Computational alanine scanning analyses of the enzyme/substrate interaction energies confirm the relevance of the P_3 , P_2 , and P_1' side chains for ligand binding. Finally, the hydrolysis of both peptides taking place at the MMP-2 active site is explored by means of hybrid quantum mechanical/molecular mechanics calculations. The computed reaction mechanisms result in rate-determining energy barriers being in consonance with the experimental k_{cat} values. Overall, the computational protocol seems to capture the subtle differences in binding and catalysis experimentally observed for the two peptide substrates. Some implications of our results for the future design of novel and more specific MMP-2 inhibitors are also discussed.

Proteins 2010; 78:1–11.
© 2009 Wiley-Liss, Inc.

Key words: enzyme catalysis; hydrolysis; metalloenzymes; molecular modelling; structural biology.

INTRODUCTION

Inhibition of matrix metalloproteinases (MMPs), a family of zinc- and calcium-dependent peptidases involved in the regulation of the cellular behavior by proteolytic processing of the extracellular environment, has long been proposed as a novel therapeutic strategy for the treatment of diseases like cancer and arthritis.^{1,2} However, the actual challenge in the design of MMP inhibitors is the development of highly specific compounds able to differentiate between the different members of the protease family.^{3–5} This requisite for specificity has been clearly established in cancer therapy, where the MMPs have been grouped as drug targets or antitargets depending on the tumorigenic effects associated to their inhibition.^{6,7}

The development of specific inhibitors for the MMPs requires the identification of subtle differences in the structure and activity of the enzymes that can be further exploited during drug design. This is a challenging task because the overall architecture of the catalytic domain is highly conserved across the MMP family. At the active site groove, the crystal structures of MMPs bound to different types of peptidomimetic inhibitors delineated the so called S_3 – S_3' binding sites, where ligands establish H-bond contacts and hydrophobic interactions with the enzyme (see Scheme 1).⁸ The specificity of these binding sites has been investigated using phage peptide libraries to characterize the substrate selection by the MMPs.^{9–14} The differences found in the kinetic parameters for the hydrolysis of the P_1 – P_1' substrate bond catalyzed by the enzymes (see Scheme 1 for peptide numbering) confirmed that a high degree of selectivity can be obtained for the individual MMPs either at binding (K_m) or catalysis (k_{cat}). However, the structural basis for such selectivity is still not well understood although it is clear that such knowledge would be useful to improve the specificity of novel MMP inhibitors.

Most likely, the design of new MMP inhibitors will require the combination of sophisticated theoretical and experimental techniques to decipher the specific structural and dynamic features of each MMP.⁴ For the MMP-2 enzyme, which is a validated drug target in cancer,⁶ we have applied different computational methods in order to answer selected questions about its structure and mode of action. First, we studied the coordination sphere of the catalytic zinc ion showing that several coordination modes are energetically

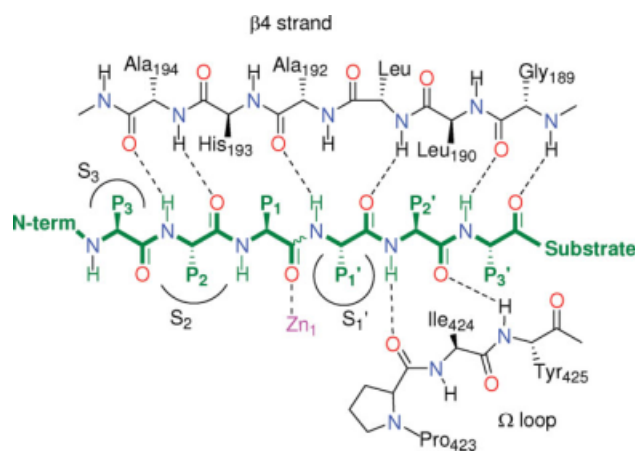
Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.

Grant sponsor: The Spanish MEC; Grant number: CTQ2004-06309.

*Correspondence to: Natalia Díaz; Departamento de Química Física y Analítica, Universidad de Oviedo, C/ Julián Clavería, 8. 33006 Oviedo (Asturias), Spain. E-mail: diaznatalia@uniovi.es.

Received 28 February 2009; Revised 11 May 2009; Accepted 25 May 2009

Published online 2 June 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22493



Scheme 1

Typical MMP-2/substrate contacts. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

accessible,¹⁵ in consonance with the variety of coordination environments experimentally observed for this ion. Then we investigated the structural and dynamical roles played by other zinc and calcium ions that are present in the crystal structures, observing that these “structural metal ions” modulate the accessibility of important anchorage points located along the active site cleft.¹⁶ The position adopted by the N-terminal coil in the active MMP-2 enzyme has also been studied.¹⁷ Binding of a small peptide substrate, which mimics the amino acid sequence of the $\alpha 1$ chain of type I collagen, to the MMP-2 catalytic domain has been intensively investigated by means of flexible docking and molecular dynamics simulations.¹⁷ For this small peptide, we have also characterized the catalytic mechanism of hydrolysis, explaining at the molecular level various results obtained from mutagenesis and kinetic experiments.¹⁷

In this article, we pursue to interrogate the MMP-2 enzyme for its selectivity towards two particular peptide substrates, whose binding and kinetic properties have been determined experimentally in a previous work on the substrate recognition profile of the enzyme: Ace-Ser-Gly-Arg-Ser-Leu-Ser-Arg~Leu-Thr-Ala-Nme (**C9**), and Ace-Ser-Gly-Ala-Val-Arg-Trp~Leu-Leu-Thr-Ala-Nme (**A13R**).¹³ These two peptide molecules are highly selective for MMP-2 with $k_{\text{cat}}/K_{\text{M}}$ values of 1.7×10^5 and $7.0 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$, respectively. The selectivity of these peptide substrates, however, results from opposite trends given that **C9** is selected by its faster reaction rate ($k_{\text{cat}} = 740 \text{ s}^{-1}$, $K_{\text{M}} = 4.4 \text{ mM}$), whereas **A13R** is favored by a stronger binding ($k_{\text{cat}} = 28 \text{ s}^{-1}$, $K_{\text{M}} = 0.4 \text{ mM}$).

To provide a detailed molecular picture that can help explain the different behavior of MMP-2 towards **C9** and **A13R**, we performed extensive molecular dynamics (MD) simulations of the MMP-2/**C9** and MMP-2/**A13R** complexes in order to characterize the enzyme/substrate

interactions. In addition, the structure of the free form of the peptide molecules in solution was also determined by computing long (100 ns) MD trajectories. From these trajectories, we performed molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations that allowed us to compute relative binding free energies. Subsequently, peptide hydrolysis occurring at the active site of the MMP-2 enzyme in complex with **C9** and **A13R** was reinvestigated by means of quantum mechanical/molecular mechanics (QM/MM) geometry optimizations. We will see that both the binding and kinetic preferences shown by MMP-2 are well reproduced by our theoretical results, providing thus a detailed molecular rationalization. The relevance of these results for designing specific inhibitors will also be discussed.

MATERIALS AND METHODS

Setup of the MMP-2/peptide complexes and flexible docking calculations

Starting coordinates for the MMP-2/**C9** and MMP-2/**A13R** complexes were obtained from our previous MD simulation of the complex formed between the MMP-2 enzyme and a peptide substrate with a collagen-like sequence (Ace-Gly-Pro-Gln-Gly~Ile-Ala-Gly-Gln-Nme) that is universally recognized by all the MMPs.¹⁷ The last point from this MD trajectory was minimized using the same settings that were used in the simulation. Subsequently, we removed all the water molecules and counterions, and we mutated the different P_4 – P_4' residues in the peptide substrate (P_1 – P_1' is the scissile peptide bond) to the corresponding ones in the **C9** and **A13R** peptides using the LEaP program.¹⁸ The residues not included in the initial substrate (P_8 – P_5 in **C9** and P_7 – P_5 in **A13R**) were built by molecular modelling assuming an initial β -sheet conformation.

To further refine the contacts between the peptide substrates and the enzyme, as well as to perform a conformational sampling of the N-terminal end of the ligands, we used the LMOD program linked to the AMBER package.¹⁹ During the LMOD calculations, all the protein residues were fixed and only the peptide substrates were allowed to move. We used the AMBER force field (parm94) coupled with the Hawkins-Cramer-Truhlar (HCT) pairwise Generalized-Born (GB) model²⁰ to mimic solvent effects. A total of 6000 LMOD iterations were computed by exploring three low-frequency vibrational modes. Eigenvectors were recalculated every 25 LMOD iterations. The LMOD calculations generated a total of 50 low energy structures for the MMP-2/**C9** and MMP-2/**A13R** complexes. Inspection of these structures confirmed that the peptide substrates maintain the main enzyme/ligand contacts that were present in the initial structure of MMP-2 with the collagen-like peptide sequence. The structures with the lowest LMOD energy

were then selected as initial points for the molecular dynamics simulation of the MMP-2/C9 and MMP-2/A13R complexes.

The initial structures of the MMP-2/C9 and MMP-2/A13R complexes were surrounded by a periodic box of TIP3P water molecules that extended 15 Å from the protein atoms. In addition, counterions were placed at the edges of the solvent box to neutralize the systems. This resulted in a total of 2717 protein atoms being solvated by ~11,000 water molecules. The parm94 version of the all-atom AMBER force field was used to model the system.²¹ For the calcium ions, we used the nonbonded representation proposed by Aqvist.²² For the zinc ions, we used a set of MM parameters that have been developed and tested by us in a previous work.¹⁶

Molecular dynamics simulations of the MMP-2/peptide complexes

Energy minimizations and MD simulations were carried out using the SANDER and PMEMD programs included in the AMBER 9.0 suite of programs.¹⁸ The solvent molecules and counterions were initially relaxed by means of energy minimizations and 50 ps of MD. Then the full systems were minimized to remove bad contacts in the initial geometry and heated gradually to 300 K during 60 ps of MD. The SHAKE algorithm was used to constraint all R—H bonds, and periodic boundary conditions were applied to simulate a continuous system. A nonbonded cutoff of 10.0 Å was used, whereas the Particle-Mesh-Ewald (PME) method was used to include the contributions of long-range interactions. The pressure (1 atm) and the temperature (300 K) of the system were controlled during the MD simulations by Berendsen's method. A 25 ns trajectory was computed for each model with a time step of 2 fs.

Coordinates were saved for analysis every 1 ps. Only the last 20.0 ns of each trajectory were analyzed using the CARNAL module of AMBER and some other specific software developed locally. 400 snapshots from each MD trajectory were used to cluster the coordinates of the peptide substrates using the MMTSB-tools.²³ The mutual similarity algorithm was used by selecting a fixed cluster radius (30°) and considering only the backbone torsion angles of the peptide. The structure in each cluster with the lowest deviation is taken as the cluster representative.

Simulations of the isolated peptides in aqueous solutions

We initially built the C9 and A13R peptide molecules by molecular modeling in a completely extended conformation. Subsequently, conformational search calculations were performed for these systems using the LMOD program and the parm94 force field coupled with the HCT GB model. The LMOD calculations generated a total of 50 low energy structures for each of the C9 and A13R

peptides. Then we selected the conformations with the lowest LMOD energy as the initial point for a 100 ns MD simulation in explicit solvent that used the same settings as those previously described for the MMP-2 complexes. The analysis of the radius of gyration of the peptides confirmed that both systems adopted a stable conformation during the simulations.

MM-PBSA calculations

Taking into account that the structural differences among the MMP/peptide complexes are relatively large, free energy calculations cannot be done with the rigorous Free Energy Perturbation approach because they would be rather difficult to converge and prohibitive in terms of computational cost. As an alternative, we use the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) approach that can perform many classes of free energy calculations.^{20,24} Basically, MM-PBSA calculations predict mean values of free energies as estimated over a series of representative snapshots extracted from classical MD simulations. The snapshots are postprocessed by removing solvent and counterion molecules. Then one estimates the free energy of one snapshot according to the following equation:

$$G_{\text{MM-PBSA}} = E_{\text{MM}} + \Delta G_{\text{solv}}^{\text{PBSA}} = E_{\text{MM}} + \Delta G_{\text{solv}}^{\text{PB}} + \gamma A \quad (1)$$

where E_{MM} is the molecular mechanics energy, and $\Delta G_{\text{solv}}^{\text{PB}}$ is the electrostatic solvation energy obtained from Poisson-Boltzmann calculations complemented with a molecular surface area term (γA) that accounts for the nonpolar part of solvation. The AMBER force field was used to compute (no cutoff) the E_{MM} terms defined in Eq. (1) using 400 snapshots extracted from the last 20 ns of the corresponding trajectory. The electrostatic contributions to the solvation free energy were determined with the Poisson-Boltzmann approach, which represents the solute as a low dielectric continuum (a value of $\epsilon_{\text{int}} = 1$ was used in the calculations) with embedded charges and the solvent as a high dielectric continuum ($\epsilon_{\text{out}} = 80$) with no salt. Atomic charges and radii were taken from the AMBER representation of the MMP-2 models. The dielectric boundary is the contact surface between the radii of the solute and the radius (1.4 Å) of a water probe molecule. The PBSA program included in the AMBER 9.0 package¹⁸ was used to solve the linearized PB equation on a cubic lattice with a grid spacing of 0.5 Å. The nonpolar solvation term was estimated by a solvent-accessible surface area (SASA) dependent term with a surface tension proportionality constant $\gamma = 72 \text{ cal mol}^{-1} \text{ \AA}^{-2}$.

The resulting average values of $G_{\text{MM-PBSA}}$ from different systems can be combined in order to estimate the corresponding free energy change. In the case of protein/ligand complexes, the $G_{\text{MM-PBSA}}$ values are usually evaluated using the snapshots from a single MD trajectory of the complex

(the one trajectory approximation) in order to minimize the statistical uncertainty of the mean value of $\Delta G_{\text{MM-PBSA}}$.²⁰ Within the one trajectory approximation, we can further analyze the $\Delta G_{\text{MM-PBSA}}$ free energies in the MMP/peptide complexes by means of a variant of the MM-PBSA protocol that is called computational alanine scanning.^{25,26} This technique allows an estimate of the contribution of an individual side chain to the overall $\Delta \Delta G_{\text{MM-PBSA}}$ free energies. To this end, alanine mutant structures are generated based on the structures of the collected MD snapshots and the MM parameters for the mutated residue are accordingly replaced with the alanine ones.

In addition to the one trajectory MM-PBSA calculations, we also computed the relative $\Delta \Delta G_{\text{MM-PBSA}}$ values by evaluating the G terms from independent MD trajectories, that is, from the 20 ns runs for the MMP/peptide complexes, and the longer trajectories (100 ns) for the isolated peptides (using 500 snapshots extracted from the last 25 ns of the corresponding trajectory). Thus, the relative binding energy of **C9** and **A13R** with respect to the MMP-2 enzyme can be estimated considering a formal substrate exchange process (see below). In these calculations, entropic effects were also taken into account by adding a $-TS_{\text{MM-GBSA}}^{\text{norm}}$ term to the $G_{\text{MM-PBSA}}$ values, where $\overline{S}_{\text{MM-GBSA}}^{\text{norm}}$ is the solute entropy as estimated by molecular mechanics normal mode calculations carried out with the NAB package.²⁷ Before the normal mode calculations, the geometries of the systems were minimized until the root-mean-squared deviation of the elements in the gradient vector was less than 10^{-5} kcal (mol Å)⁻¹. These minimizations and the subsequent normal mode calculations were carried out using the HCT GB model for representing a solvent environment.

Conformational entropy calculations

The total configurational entropy of a solute molecule can be estimated from a MD simulation by means of the following approximation²⁸:

$$S_{\text{total}} = \overline{S}_{\text{vib}} + S_{\text{conf}} \quad (2)$$

where $\overline{S}_{\text{vib}}$ is the average vibrational entropy derived from the $\overline{S}_{\text{MM-GBSA}}^{\text{norm}}$ values, and S_{conf} is the conformational entropy of the whole MD trajectory, which, in turn, can be computed using the Shannon entropy or information entropy:

$$S_{\text{conf}} = -R \sum_{\alpha} p_{\alpha} \ln p_{\alpha} \quad (3)$$

where the index α runs over all possible conformers of the solute molecule and p_{α} is the statistical weight of the α -conformer. Each conformer can be labeled univocally with an array $\{A_i\}_i = 1, N$ where N is the number of torsion angles of the molecule and A_i is an index that defines the conformational state (e.g., $g+$, $g-$, $anti$) of

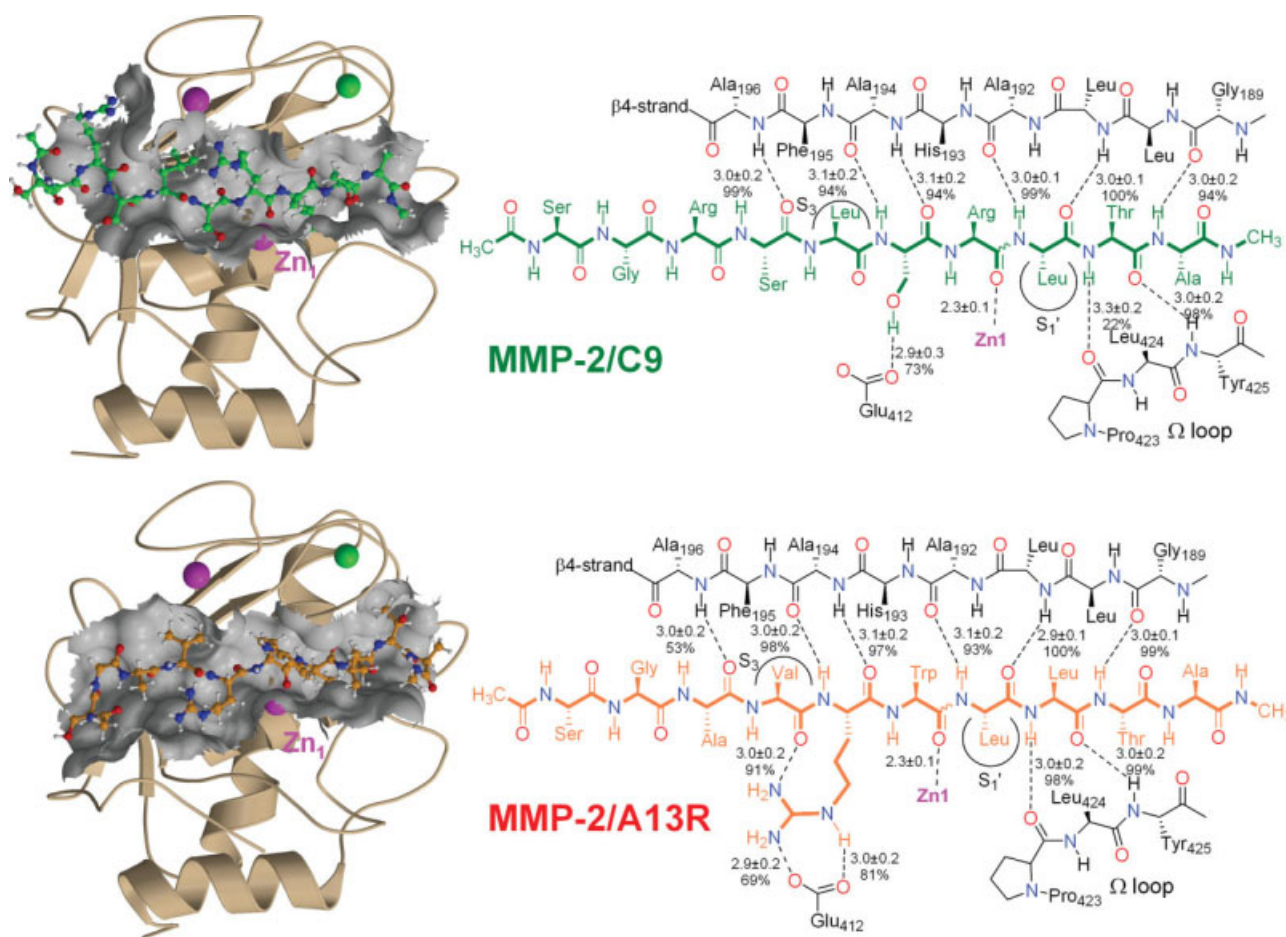
the i -th torsion angle. Based on the converged probability distributions of the individual torsion angles, it is possible to estimate the conformational entropy of a solute molecule by expanding the global informational entropy of Eq. (3) in terms of the so-called Mutual Information Functions (MIF).²⁹ However, to make the S_{conf} calculations feasible, it is necessary to reduce the dimensionality of the problem as well as to truncate the expansion of the MIFs. Taking into account that the largest conformational changes occurring upon peptide binding correspond to their backbone chains, we computed only the S_{conf} values arising from the conformational freedom of the ϕ and ψ torsion angles of **C9** and **A13R**. In addition, the expansion of MIFs was truncated to fourth-order as suggested by former test calculations on peptide systems.³⁰ For the **C9** and **A13R** systems, their S_{conf} values were estimated by using the coordinates from the last 75/20 ns of the simulation of the free/bound peptides, respectively, and following similar prescriptions to those described elsewhere.³⁰

QM/MM calculations of the reaction mechanism

From the 25 ns MD simulations of the MMP-2/**C9** and MMP-2/**A13R** complexes, we selected representative snapshots to investigate the reaction mechanism by means of QM/MM calculations. The selected systems were truncated by including the whole catalytic domain of MMP-2 with two Ca²⁺ and two Zn²⁺ ions bound to the enzyme, the corresponding peptide substrate (**C9** or **A13R**), and the 1000 water molecules closest to the Zn₁ site.

QM/MM geometry optimizations were carried out using the version 6.0 of the QSite program.³¹ QM-MM interfaces were placed at the His₄₀₃-C α -N, Phe₄₀₅-N-C α , His₄₀₇-C β -C α , and His₄₁₃-C β -C α bonds of MMP-2 (residue numbering as in the 1CK7 structure³²) and the P₂-C-C α and the P₂'-N-C α bonds of the peptide substrates **C9** and **A13R**. In addition, the QM region also included the catalytic Zn₁ ion and Wat₁. This partitioning resulted in a QM region comprising a total of 105 QM atoms with a net charge of +2 for **C9** and +1 for **A13R** that was described at the B3LYP/LACVP* level of theory (959 and 995 basis functions, respectively).³³ The rest of the protein and solvent atoms were treated with the OPLS-AA force field.³⁴ This mixing of the B3LYP/LACVP* level of theory with the OPLS force field has been specifically optimized for modeling protein active sites.³⁵

During QM/MM geometry optimizations, all the protein residues that contact the QM region were allowed to move. These included Phe₁₁₃, Tyr₁₈₂, Gly₁₈₉-Ala₁₉₆, Val₄₀₀, His₄₀₃-His₄₀₇, Glu₄₁₂, His₄₁₃, and Ala₄₂₂-Tyr₄₂₅. In addition, Zn₁, the whole peptide substrates and the closest water molecules were relaxed. The position of the rest of the protein and solvent molecules was frozen during the QM/MM calculations. Minimizations were performed

**Figure 1**

Schematic representation of the main enzyme/substrate binding determinants (average distances are in Å). Ribbon models and molecular surface representations of the last snapshot from the MMP-2/C9 and MMP-2/A13R trajectories (the C9 and A13 peptides are depicted in ball-and-sticks with carbon atoms shown in green and orange, respectively).

with no cutoff until the root-mean-squared residual gradient was less than 5.0×10^{-4} in au, permitting thus a proper relaxation of the QM region and the proximal protein and solvent residues.

Long-range solute–solvent electrostatic interactions were estimated by means of PB calculations after having deleted the coordinates of the majority of the MM water molecules in the optimized QM/MM structures (only the 50 water molecules that are closer to the Zn₁ atom were kept). Single-point QM/MM calculations were then carried out on the same structures to give the $E_{\text{QM/MM}}$ energies. The PB solvation free energy ($\Delta G_{\text{solv}}^{\text{PB}}$) was determined using the Delphi program.³⁶ The protein was represented as a low dielectric continuum (a value of $\epsilon_{\text{int}} = 1$ was used in the calculations) with embedded charges and the solvent as a high dielectric continuum ($\epsilon_{\text{out}} = 80$) with no salt. The OPLS-AA atomic charges were used for the protein atoms excepting those atoms that were within the QM region during QM/MM geometry optimizations. For these atoms, we used their ESP

charges derived from single-point B3LYP/LACVP* calculations on the QM region of the corresponding QM/MM optimized geometries after having placed H-link atoms at the QM/MM cuts (a zero value to the atomic charges of the H-link atoms was assigned in the ESP fitting procedure using the RESP program¹⁸).

RESULTS AND DISCUSSION

MD simulations of the Michaelis complexes

Binding of the C9 and A13R peptides to the MMP-2 active site does not induce significant changes in the global structure of the host enzyme. For instance, the root mean squared deviation (RMSD) with respect to the X-ray structure, and the root mean squared flexibility (RMSF) with respect the mean structure, are comparable in both simulations and remain similar to those previously reported¹⁷ for the native enzyme (see Supporting Information Table SI). The radius of gyration and the solvent accessible

surface area (SASA) of the enzyme are also quite similar in the two trajectories and in previous simulations. Thus, we conclude that the active site cleft is well preorganized to interact with different peptide sequences.

In Figure 1, we see that, in the MMP-2 active site, the two peptide substrates, **C9** and **A13R**, adopt an extended conformation around the scissile peptide bond. Note, however, that **C9** is not symmetrically placed in the MMP-2 active site cleft: only the last three aminoacids (-Leu-Thr-Ala-Nme in Fig. 1) interact with the primed subsites S_1' - S_3' while the rest of the **C9** chain is oriented towards the nonprimed ones. The first three residues of **C9** at the N-terminal end, which do not give stable contacts with any MMP-2 residue, are structurally disordered along the MD simulation. In contrast, the MMP-2 bound **A13R** peptide is placed more symmetrically along the active site region and, according to the RMSF values in Supporting Information Table SII, it is more rigid during the MD simulation (the backbone RMSFs are 1.27 ± 0.37 and 0.98 ± 0.32 Å for **C9** and **A13R**, respectively). Nevertheless, the P_3 - P_3' backbone atoms of **C9** and **A13R** give very similar interactions with the MMP-2 residues from the β_4 strand (Gly₁₈₉, Leu₁₉₁, Ala₁₉₂, Ala₁₉₄, Ala₁₉₆) and from the Ω -loop (Pro₄₂₃ and Tyr₄₂₅) with average distances of ~ 3 Å and 90–100% of occupancy. In addition, the P_1 carbonyl group of the scissile peptide bonds interacts with the Zn₁ ion with average distances of around 2.3 ± 0.1 Å (we note that a Zn...O- P_1 bond is not explicitly defined in the force field representation of Zn₁). The largest differences between the enzyme/substrate backbone contacts of both complexes arise at the Ala₁₉₆-NH...O- P_4 and Pro₄₂₄-O...HN- P_2' interactions (see Fig. 1). Concerning the peptide side chains, the polar residue at P_2 (Ser for **C9** and Arg for **A13R**) inter-

acts with the carboxylate group of Glu₄₁₂, whereas the hydrophobic P_3 and P_1' residues (Leu/Leu for **C9** and Val/Leu for **A13R**) remain bound in the S_3 and S_1' cavities. In addition, the P_1 tryptophan residue in **A13R** gives a hydrophobic contact with the Leu₁₉₀ side chain (4.90 ± 0.59 Å), which results in a lower flexibility for this residue as compared with the P_1 arginine residue in **C9**. For example, the RMSF values for the side chain atoms in Trp- P_1 (**A13R**) and Arg- P_1 (**C9**) are 0.41 ± 0.01 and 0.66 ± 0.08 Å, respectively. The P_3' alanine in **C9** interacts with the hydrophobic Leu₁₉₁ (5.13 ± 0.88 Å) while the same position in **A13R**, which is occupied by a threonine residue, contacts the Leu₁₉₁ (4.62 ± 0.29 Å) and the Asp₁₈₈ (66% 2.86 ± 0.26 Å) side chains.

Globally, all the MMP-2/substrate contacts contribute to anchor both peptide substrates within the MMP-2 active site and, simultaneously, result in a relative good orientation between the scissile peptide bond of the substrate and the zinc-bound nucleophilic water molecule. Thus, the average values for the Wat₁-O...C- P_1 distance (2.9 ± 0.1 Å for both substrates) and the Wat₁-O...C- P_1 ...O- P_1 angle (75 ± 5 and $72 \pm 5^\circ$ for **C9** and **A13R**, respectively) further confirm that the two peptide substrates adopt a reactive configuration in the MMP-2 active site and that the Zn₁-bound water molecule is properly oriented for nucleophilic attack.

MD simulations of the free peptide substrates

The problem of finding out the structure of the isolated forms of the **C9** and **A13R** decapeptides can be seen as a polypeptide folding problem which, in turn, can be solved computationally for small peptides (6–15 residues) through extensive MD samplings.³⁷ As men-

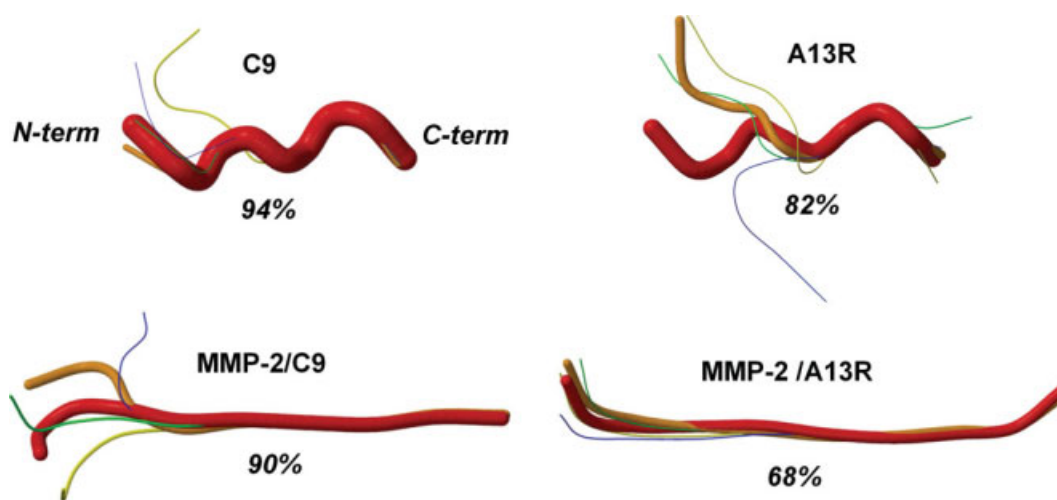


Figure 2

Superposition of the most populated representative structures derived from clustering analyses for the **C9** and **A13R** peptides both in their free and enzyme-bound states. Thickness of the models corresponds to the number of snapshots represented by each model.

Table I

Average Values for the (one trajectory) MM-PBSA Energy Components (kcal mol⁻¹) of the MMP-2 Enzyme in Complex with **C9** and **A13R**. Standard Errors of the Mean Values are Indicated in Parentheses. The Changes in the Average Values Due to Selected Computational Alanine Mutations Located at the Peptide Substrate are Also Indicated

	$\Delta\bar{E}_{MM}$	$\Delta\bar{E}_{elec}$	$\Delta\bar{E}_{vdW}$	$\Delta\bar{G}_{solv}^{PBSA}$	$\Delta\bar{G}_{int}$
MMP-2/C9	-265.2 (1.2)	-181.3 (1.3)	-84.0 (0.3)	198.0 (1.1)	-67.2 (0.3)
<i>P</i> ₇ (Ser→Ala)	4.1	3.7	0.5	-4.3	-0.2
<i>P</i> ₅ (Arg→Ala)	33.7	32.5	1.3	-31.8	1.9
<i>P</i> ₄ (Ser→Ala)	1.3	0.6	0.8	-1.7	-0.4
<i>P</i> ₃ (Leu→Ala)	7.4	0.4	7.1	-1.7	5.7
<i>P</i> ₂ (Ser→Ala)	15.3	15.8	-0.4	-10.8	4.5
<i>P</i> ₁ (Arg→Ala)	56.7	53.9	2.8	-56.0	0.8
<i>P</i> ₁ ' (Leu→Ala)	9.3	0.1	9.2	-0.8	8.4
<i>P</i> ₂ ' (Thr→Ala)	4.0	2.0	2.0	-3.3	0.6
MMP-2/A13R	-269.6 (0.7)	-176.5 (0.7)	-93.2 (0.3)	191.1 (0.5)	-78.6 (0.4)
<i>P</i> ₆ (Ser→Ala)	4.8	4.8	0.2	-4.7	0.2
<i>P</i> ₃ (Val→Ala)	4.9	0.3	4.6	0.1	5.1
<i>P</i> ₂ (Arg→Ala)	78.6	74.2	4.5	-69.5	9.3
<i>P</i> ₁ (Trp→Ala)	7.8	2.2	5.7	-4.5	3.3
<i>P</i> ₁ ' (Leu→Ala)	9.4	0.1	9.3	-0.5	8.9
<i>P</i> ₂ ' (Leu→Ala)	3.5	0.3	3.3	-0.4	3.2
<i>P</i> ₃ ' (Thr→Ala)	13.4	9.8	3.7	-6.8	6.7

tioned in Methods, prior to MD simulations we carried out conformational search calculations using the LMOD algorithm. For the two peptide systems, the subsequent 100 ns MD trajectories evolved rapidly towards stable structures, with the same radius of gyration for **C9** and **A13R** fluctuating smoothly (5.9 ± 0.2 Å). Secondary structure analyses assign a helical conformation to the central (i.e., 2–8) residues in ~85% and ~44% of the analyzed snapshots for **C9** and **A13R**, respectively (see Fig. 2). The RMSF values for the backbone atoms in Supporting Information Table SII confirm that **C9** is less flexible than **A13R** in solution (0.52 ± 0.26 and 1.50 ± 0.49 Å, respectively). The first three residues of **A13R** tend to adopt a coil structure in the free state of this peptide molecule. Although the AMBER parm94 force field tends to overestimate the helical content of peptide molecules, the propensity of **C9** and **A13R** to adopt a helical conformation seems in consonance with the fact that MMP-2 is able to bind and hydrolyze collagen peptide chains that have also a helical structure. It may also be interesting to note that unwinding of a helical conformation to give a nearly extended conformation seems also a plausible mechanism for substrate binding.

MM-PBSA calculations

Clearly, the structural analyses of the enzyme/ligand contacts do not allow us to determine the binding preferences of the MMP-2 enzyme against the **C9** and **A13R** peptide substrates. For such purpose, we need to estimate their relative binding energy within the MMP-2 active site.

First, we carried out MM-PBSA calculations assuming the one trajectory approximation, that is, using only the

MD snapshots from the MMP-2/peptide complexes. In this way, the resulting $\Delta G_{MM-PBSA}$ values are better considered as interaction energies ($\Delta\bar{G}_{int}$) between the MMP-2 enzyme and the enzyme-bound peptide molecules. The results, collected in Table I, show that the **A13R** peptide has an interaction free energy that is 11.4 kcal mol⁻¹ larger in absolute value than the one corresponding to **C9**. Inspection of the free energy components shows that the van der Waals ($\Delta\bar{E}_{vdW}$) and desolvation contributions ($\Delta\bar{G}_{solv}$) favor the interaction of the **A13R** substrate with MMP-2, compensating thus the more stable electrostatic interaction ($\Delta\bar{E}_{elec}$) obtained for the (+2)-charged **C9** (see Table I). The more polar character of **C9** (Ace-Ser-Gly-Arg-Ser-Leu-Ser-Arg~Leu-Thr-Ala-Nme) when compared with **A13R** (Ace-Ser-Gly-Ala-Val-Arg-Trp~Leu-Leu-Thr-Ala-Nme) and the more evenly positioning of **A13R** over the active site cleft, explain well the differences observed in the various energetic components.

To gain further insight into the contributions of the individual side chains of the peptide substrate to the overall protein/substrate interaction energy, computational alanine scanning (CAS) analyses were performed on the same set of snapshots by mutating substrate residues to alanine and recomputing the $\Delta\bar{G}_{int}$ values. The results, also included in Table I, reveal that mutation of the hydrophobic side chain at *P*₁' decreases by 8–9 kcal mol⁻¹ the enzyme-peptide interaction energy either for **C9** or **A13R**. This result is not entirely unexpected because the *S*₁' hydrophobic pocket, together with the Zn₁ ion, are usually considered as the main binding determinants for inhibitor design. However, the CAS results also highlight other critical binding points. Particularly, an important decrease in the interaction energy (4.5 kcal mol⁻¹ for **C9** and 9.3 kcal mol⁻¹ for **A13R**) is observed after replacing the *P*₂ residue by Ala (see Table I) due to the loss of the direct contact between the *P*₂ side chain and the Glu₄₁₂ carboxylate group. Interestingly, in other MMPs, the Glu₄₁₂ residue is replaced by other polar residues, large hydrophobic side chains, and even a Gly residue. This variability in the *S*₂ site across the MMP family, together with the influence of the *P*₂ residue on the interaction energy as shown by our analysis, point to this region as a hot spot for substrate distinction that could be exploited in the design of more specific MMP inhibitors.

In terms of the *K*_M values reported for **C9** (4.4 mM) and **A13R** (0.4 mM),¹³ and assuming that *K*_M coincides with the equilibrium constant for the dissociation of the MMP-2/peptide complexes, it turns out that MMP-2 binds preferentially to the **A13R** peptide. From the *K*_M values at 37°C, the relative $\Delta G_{binding}$ of MMP-2/**A13R** with respect to MMP-2/**C9** amounts to -1.5 kcal mol⁻¹. Although this experimental observation is in consonance with the one trajectory MM-PBSA calculations and the related CAS analyses, it is also clear that basing solely on the relative MM-PBSA interaction energies ($\Delta\Delta\bar{G}_{int}$)

Table II

Average Values for the (Independent Trajectories) MM-PBSA Energy Components (kcal mol⁻¹) of the MMP-2 Enzyme in Complex with C9 and A13R and the Free Peptide Substrates. Standard Errors of the Mean Values are Indicated in Parentheses

	\bar{E}_{MM}^a	$\Delta\bar{G}_{\text{sol}}^a$	$\bar{S}_{\text{MM-GBSA}}^{\text{norm}b}$	$S_{\text{conf}}^{b,c}$	$\bar{G}^{a,d}$
MMP-2/C9	-3533.0 (4.6)	-1969.1 (3.7)	6476.2 (2.0)	7.2	-7445.3 (2.1)
MMP-2/A13R	-3467.0 (3.7)	-1848.1 (3.1)	6451.7 (1.4)	5.7	-7249.8 (2.0)
C9	-256.7 (0.9)	-214.1 (0.7)	469.3 (0.5)	5.2	-611.6 (0.5)
A13R	-154.1 (0.6)	-115.4 (0.4)	470.0 (0.3)	10.9	-410.5 (0.5)
MMP-2/C9 + A13R → MMP-2/A13R + C9					
	-36.6 (6.0)	22.3 (4.9)	-25.9 (2.5)	-7.2	-3.4 (2.9)

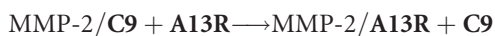
^aIn kcal/mol.

^bIn cal/(mol K).

^cConformational entropy due to the backbone torsion angles of the C9 and A13R peptides.

^d $\bar{G} = \bar{E}_{\text{MM}} + \Delta\bar{G}_{\text{sol}} - T(\bar{S}_{\text{MM-GBSA}}^{\text{norm}}) + S_{\text{conf}}$.

-11.4 kcal mol⁻¹) we overestimate the relative stability of the MMP-2/A13R complex. Therefore, to obtain a better assessment of the relative A13R and C9 binding strength, we computed the free energy change ($\Delta\Delta G_{\text{binding}}$) for the following peptide exchange process:



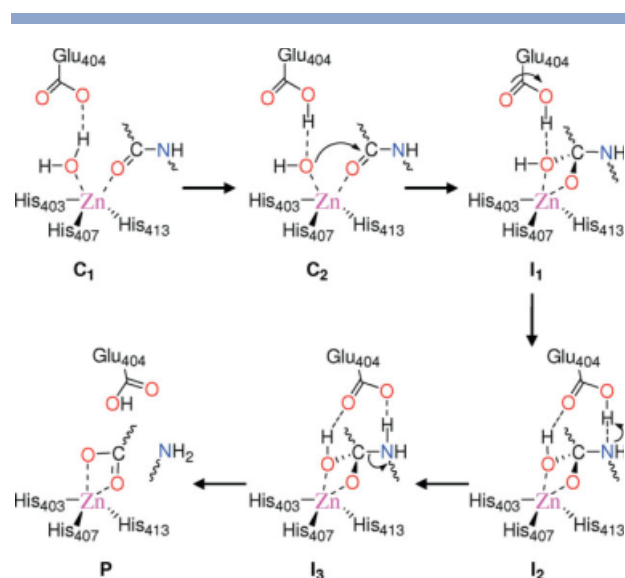
using the MM-PBSA free energies derived from the independent MD trajectories and including solute entropic effects. Besides weighting the direct enzyme-substrate interactions, the corresponding $\Delta\Delta G_{\text{binding}}$ values reflect the enthalpic and entropic contributions due to peptide reorganization upon substrate binding, which can be particularly important because the enzyme-bound C9 and A13R peptides adopt a more extended conformation than in their free state in aqueous solution (see Fig. 2). As a matter of fact, we see in Table II that all the energetic terms contributing to $\Delta\Delta G_{\text{binding}}$ vary significantly during the exchange process. The gas-phase MM energy (-37 kcal mol⁻¹) favors strongly the binding of A13R vs. C9, but this effect is largely compensated by the relative changes in the solvation free energy (+22 kcal mol⁻¹) and in the normal mode solute entropy (+8 kcal mol⁻¹ at 300 K). Similarly, the conformational entropy associated to the backbone torsional motions of the peptide molecules also disfavors the binding of the A13R peptide by ~2 kcal mol⁻¹. Curiously, the backbone conformational entropy of C9 is larger in its MMP-2 bound form than in its free state (see Table II), in consonance with the RMSF values in Supporting Information Table SII and the clustering analyses reported in Figure 2. This effect is related with the unsymmetrical placement of C9 in the active site cleft, where the C9 residues Ser(P₇), Gly(P₆), and Arg(P₅) do not establish stable interactions

with the enzyme and become quite mobile, and with the high stability of the α -helix structure observed for the C9 unbound state. In contrast, the hydrophobic A13R peptide, which is more flexible in solution than C9, fits better to the active site and gives a more rigid MMP-2/peptide complex.

The overall $\Delta\Delta G_{\text{binding}}$ energy for the substrate exchange process is still negative, -3.4 kcal mol⁻¹, showing thus that the MMP-2 enzyme prefers to bind A13R over C9. Although its statistical uncertainty (2.9 kcal mol⁻¹) does not allow us to make a clear-cut prediction, it is clear that the computed average value for $\Delta\Delta G_{\text{binding}}$ (-3.4 kcal mol⁻¹) shows a more reasonable agreement with the energy difference derived from the experimentally-reported K_M values (-1.5 kcal mol⁻¹) than the values estimated from the one trajectory MM-PBSA interaction energies (-11.4).

Reaction mechanism

According to our QM/MM calculations, the hydrolysis of the C9 and A13R peptides catalyzed by MMP-2 proceeds through the same molecular mechanism to that previously found for the hydrolysis of a collagen-like peptide sequence.³⁸ In this mechanism, which is similar to the traditionally accepted mechanism for thermolysin, the Zn₁-bound water molecule is activated by the conserved Glu₄₀₄ residue and the resulting Zn-OH moiety attacks the P₁ carbonyl group to give a tetrahedral intermediate (see Scheme 2). Following an H-bond rearrangement process involving the protonated Glu₄₀₄ residue, a Glu₄₀₄-COOH → N-P₁' proton transfer takes place, which in turn triggers the breaking of the substrate C-N bond.



Scheme 2

Schematic representation of the QM/MM reaction mechanism. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

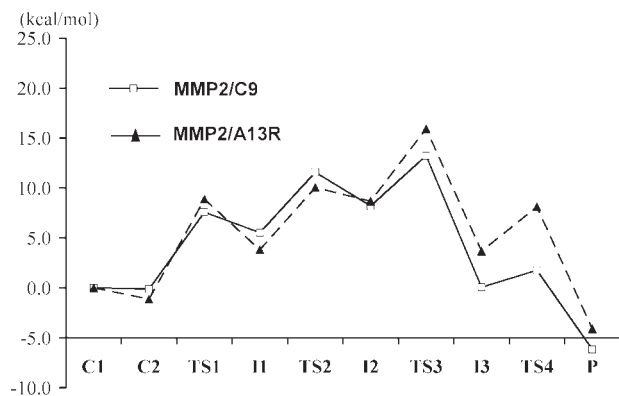


Figure 3

QM/MM energy profiles (in kcal/mol) for the hydrolysis of **C9** and **A13R** in the active site of the MMP-2 enzyme.

Figure 3 represents schematically the energy profiles for the hydrolysis of the **C9** and **A13R** peptides as estimated by our QM/MM and Poisson-Boltzmann calculations. All the critical structures located along the QM/MM pathways are displayed in Supporting Information Figures S1 and S2 and the main geometrical information for the two profiles is collected in Supporting Information Tables SIII and SIV. Relative energies are reported also in the Supporting Information (Table SV).

As in our previous study,³⁸ two prereactive complexes characterized by a $[\text{Zn}-\text{OH}_2]^{2+} \dots \text{OOC}-\text{Glu}_{404}$ (**C**₁) and a $[\text{Zn}-\text{OH}]^+ \dots \text{HOOC}-\text{Glu}_{404}$ (**C**₂) associations, respectively, were located on the QM/MM potential energy surface for **C9** and **A13R**. These two critical points are extremely similar both in their structure and in stability, although **C**₂ is slightly more stable than **C**₁ by 0.1 kcal mol⁻¹ and 1.1 kcal mol⁻¹ for **C9** and **A13R** (see Fig. 3). The formation of the first tetrahedral intermediate (**I**₁) occurs through a low energy barrier transition structure (**TS**₁). In the second step, the N-*P*₁' atom becomes pyramidalized and the H-bond rearrangement required for Glu₄₀₄ to act as a proton shuttle takes place through the **TS**₂ structure, which is 11.7 and 11.2 kcal mol⁻¹ above **C**₂ for **C9** and **A13R**, respectively. The resulting tetrahedral intermediate, **I**₂, is close in structure and energy to the previous **TS**₂, and is connected with a third intermediate structure (**I**₃) through a Glu₄₀₄-Oε2H→:N-*P*₁' proton transfer transition structure (**TS**₃), which turns out to be the rate-determining TS (see below). At **I**₃, the reactive C—N bond is weakened (1.6–1.7 Å) and further elongation of this bond in the **TS**₄ structure leads to the product complex **P**, in which the Zn₁ ion is coordinated in a bidentate manner by the resulting carboxylate group and the Glu₄₀₄ side chain is neutralized.

Comparison of the different critical points obtained for the hydrolysis of **C9** and **A13R** shows that Zn₁ main-

tains a similar five-coordination environment in the two reaction profiles (see Supporting Information Table SIII). Concerning the reactive events, the equivalent critical structures in the two reaction pathways present comparable bond distances and bond angles involving the reactive atoms. For example, the C...O distance for nucleophilic attack at **TS**₁ is ~2.0 Å in both systems and the Glu₄₀₄-Oε2...H and H...N-*P*₁' distances involved in the proton transfer at **TS**₃ amount to ~1.2 and ~1.4 Å, respectively, both for **C9** and **A13R**. The main differences arise at the transition structures for the H-bond rearrangement (**TS**₂), which is more advanced for **C9** (e.g., the Wat₁-H1...O-*P*₂ and Wat₁-H1...Oε1-Glu₄₀₄ distances at **TS**₂ are 2.5/2.2 and 1.9/2.2 Å for **C9/A13R**).

The energy barriers with respect to the most stable prereactive complex (**C**₂) are 13.4 kcal mol⁻¹ for **C9** and 17.1 kcal mol⁻¹ for **A13R**. Although neither the thermal contributions to free energy from the reacting QM atoms or the free energy corrections due to structural fluctuations of the protein and solvent environment have been included in the present QM/MM calculations, we note that these contributions would not be large and cancel partially to each other according to our previous study.³⁸ Hence, the relatively large difference (3.7 kcal mol⁻¹) observed in the rate-determining QM/MM energy barriers allows us to safely predict that **C9** would react faster than **A13R**, what is in agreement with experimental observations. Moreover, the computed energy barriers are close to the activation free energies derived from experimental *k*_{cat} data using the classical transition state formula, which are 14.2 kcal mol⁻¹ for **C9** and 16.2 kcal mol⁻¹ for **A13R** at 37°C.

Besides reproducing the MMP-2 kinetic preference for **C9**, the QM/MM calculations can give insight into the origin of this effect. To this end, we carried out single-point QM calculations using only the coordinates of the “QM region” complemented with H-link atoms (see Supporting Information Table SV). For the hydrolysis of **A13R**, in which a bulky non-polar Trp residue at the *P*₁ site is fully included within the QM region, we found that the gas-phase QM and QM/MM energy profiles are quite similar, what suggests that the hydrolysis of **A13R** is quite insensitive to environmental effects. In contrast, for the **C9** peptide, the positively charged side chain of the Arg residue located at the *P*₁ position, which is also relatively close to the Zn₁ site (the Zn₁...Nε-Arg distance lies within the 7.4–7.6 Å range), reinforces substantially the intrinsic stability of the QM region along the reaction coordinate. In this case, the **C9** transition structures and intermediates, in which the developing negative charge at the carbonylic O-*P*₁ atom is oriented towards the Arg-*P*₁ side chain, have very low gas-phase QM relative energies, but correspondingly they have a worse “solvation” stabilization in their interaction with the protein and solvent surroundings than in the initial prereactive complexes. Thus, the rate enhancement due to the Arg substitution

at P_1 is determined by the balance between near field electrostatic interactions (described by the gas-phase QM energies) and far field electrostatic effects (described by the QM/MM + PB solvation energies).

SUMMARY AND CONCLUSIONS

To find out to what extent current simulation methodologies can describe (at least) semiquantitatively some fine tuning effects controlling the interaction of the MMP-2 enzyme with different types of substrates and inhibitors, we have analyzed in this work the experimental trends found in the K_M and k_{cat} values for the **A13R** and **C9** pair of peptide substrates, showing that they can be reproduced by combining extensive MD samplings followed by MM-PBSA and QM/MM calculations.

On one hand, our MM-PBSA calculations on the MD snapshots from independent trajectories complemented with an estimation of the conformational entropy of the peptide molecules predict that the MMP-2 enzyme binds preferentially to the **A13R** peptide by ~ 3 kcal mol⁻¹. The analysis of the various free energy contributions reveals that the actual binding selectivity for one or another peptide is a mixture of enthalpic, solvation, and entropic effects that must be described in a balanced manner. In fact, some of these effects cannot be captured by the MM-PBSA calculations using the one trajectory approximation, which in our problem tend to overestimate the binding preference for **A13R**. Nonetheless, it is interesting to note that the computational alanine scanning variant of the MM-PBSA approach confirms the relevance of the P_3 , P_2 , and P_1' sites for enzyme-substrate interaction and point to the nonprimed sites, specially to the non-conserved P_2 site, as relevant points for inhibitor design capable of increasing the specificity of the compounds across the MMP family.

On the other hand, the QM/MM energy profiles for the hydrolysis of the two peptide substrates in the MMP-2 active site confirm that the hydrolytic degradation of the MMP-2/**C9** Michaelis complex is faster than that of MMP-2/**A13R**. This kinetic preference hardly modifies the internal geometry of the critical structures along the reaction coordinate given that the same number of TSs and intermediates with very similar structural properties are located for **C9** and **A13R**. Our calculations suggest that the Arg side chain at the P_1 site in **C9** help stabilize all the reactive configurations through electrostatic interactions, although this effect is partially dampened by long-range interactions in the solvated enzyme. The rate-determining energy barriers (~ 13 and ~ 17 kcal mol⁻¹ for **C9** and **A13R**), which are in consonance with the experimental k_{cat} values, correspond to the protonation of the leaving amino group. For this process to occur, the N- P_1' atom of the substrate approaches to the protonated Glu₄₀₄ side chain, but this shift is partially impeded due to steric interactions between

the substrate and the primed binding subsites. This observation may suggest that bulky groups interacting with the primed region of the active site of the enzyme should be avoided in the design of mechanism-based inhibitors.

Effective inhibitors of the MMPs require K_I values in the low nM range. Usually, these values result from the combination of a potent Zn-binding group, frequently an hydroxamate, and a hydrophobic side chain that binds within the S_1' cavity of the enzyme. However, this combination of functional groups drastically limits the opportunities to increase the desired specificity of the inhibitors against different MMP enzymes. Thus, the new paradigm in inhibitor design looks for weaker Zn-chelating groups at the cost of exploring additional binding sites to achieve a good binding energy.⁷ In principle, the more binding sites explored by the ligands, the more opportunities to increase the specificity of the compounds. From the analysis of the MMP-2/peptide complexes examined in this work, we have characterized a number of interactions at the substrate backbone (P_2 , P_1' , and P_2') and side chains (P_3 , P_2 , and P_1') that significantly contribute to the interaction free energies. Particularly, the inclusion of a positively charged residue capable of interacting with Glu₄₁₂ at the S_2 site could contribute to increase the specificity towards MMP-2. Finally, it should also be noted that the computational design of inhibitors able to contact simultaneously all the relevant S_2 - S_2' binding sites will require to carefully evaluate the complex changes in the enthalpic and entropic contributions upon enzyme/inhibitor binding.

ACKNOWLEDGMENTS

E.S. and N.D. thank MEC for their FPU and Ramon y Cajal contracts, respectively. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Barcelona Supercomputing Center—Centro Nacional de Supercomputación.

REFERENCES

1. Sternlicht MD, Werb Z. How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* 2001;17:463–516.
2. McCawley LJ, Matrisian LM. Matrix metalloproteinases: they're not just for matrix anymore! *Curr Opin Cell Biol* 2001;13:534–540.
3. Cuniasse P, Devel L, Makaritis A, Beau F, Georgiadis D, Matziari M, Yiotakis A, Dive V. Future challenges facing the development of specific active-site-directed synthetic inhibitors of MMPs. *Biochimie* 2005;87:393–402.
4. Rao BG. Recent developments in the design of specific matrix metalloproteinase inhibitors aided by structural and computational studies. *Curr Pharm Des* 2005;11:295–322.
5. Nuti E, Tuccinardi T, Rossello A. Matrix metalloproteinase inhibitors: new challenges in the era of post broad-spectrum inhibitors. *Curr Pharm Des* 2007;13:2087–2100.
6. Overall CM, Kleifeld O. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nat Rev Cancer* 2006;6:227–239.

7. Overall CM, Kleifeld O. Towards third generation matrix metalloproteinase inhibitors for cancer therapy. *Brit J Cancer* 2006; 94:941–946.
8. Maskos K. Crystal structures of MMPs in complex with physiological and pharmacological inhibitors. *Biochimie* 2005;87:249–263.
9. Netzel-Arnett S, Sang Q-X, Moore WGI, Navre M, Birkedal-Hansen H, Van Wart HE. Comparative sequences specificities of human 72- and 92-kDa gelatinases (type IV collagenases) and PUMP (matrilysin). *Biochemistry* 1993;32:6427–6432.
10. Smith MM, Shi L, Navre M. Rapid identification of highly active and selective substrates for stromelysin and matrilysin using bacteriophage peptide display libraries. *J Biol Chem* 1995;270:6440–6449.
11. Turk BE, Huang LL, Piro ET, Cantley LC. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat Biotech* 2001;19:661–667.
12. Kridel SJ, Chen E, Kotra LP, Howard EW, Mobashery S, Smith JW. Substrate hydrolysis by matrix metalloproteinase-9. *J Biol Chem* 2001;276:20572–20578.
13. Chen EI, Kridel SJ, Howard EW, Li W, Godzik A, Smith JW. A unique substrate recognition profile for matrix metalloproteinase-2. *J Biol Chem* 2002;277:4485–4491.
14. Chen EI, Li W, Godzik A, Howard EW, Smith JW. A residue in the S2 subsite controls substrate selectivity of matrix metalloproteinase-2 and matrix metalloproteinase-9. *J Biol Chem* 2003;278:17158–17163.
15. Díaz N, Suárez D, Sordo TL. Quantum chemical study on the coordination environment of the catalytic zinc ion in matrix metalloproteinases. *J Phys Chem B* 2006;110:24222–24230.
16. Díaz N, Suárez D. Molecular dynamics simulations of matrix metalloproteinase 2: the role of the structural metal ions. *Biochemistry* 2007;46:8943–8952.
17. Díaz N, Suárez D. Molecular dynamics simulations of the active matrix metalloproteinase-2: positioning of the N-terminal fragment and binding of a small peptide substrate. *Proteins: Struct Funct Bioinf* 2008;72:50–61.
18. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Matthews DH, Schafmeister C, Ross WS, Kollman PA. AMBER 9. San Francisco: University of California; 2006.
19. Kolossváry I, Guida WC. Low mode search. An efficient, automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides *J Am Chem Soc* 1996;118:5011–5019.
20. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* 2000;33:889–897.
21. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 1995;117:5179–5197.
22. Aqvist J. Ion-water interaction potentials derived from free energy perturbation simulations. *J Phys Chem* 1990;94:8021–8024.
23. Feig M, Karanicolas J, Brooks CL. MMTSB tool set: Enhanced sampling and multiscale modelling methods for applications in structural biology. *J Mol Graph Model* 2004;22:377–395.
24. Gohlke H, Case DA. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem* 2003;25:238–250.
25. Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J Comput Chem* 2002;23:15–27.
26. Massova I, Kollman PA. Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J Am Chem Soc* 1999;121:8133–8143.
27. Macke T, Case DA. Modeling unusual nucleic acid structures. In: Leontes NB, SantaLucia JJ, editors. *Molecular modeling of nucleic acids*. Washington, DC: American Chemical Society; 1998. pp 379–393.
28. Karplus M, Ichiye T, Pettit BM. Configurational entropy of native proteins. *Biophys J* 1987;52:1083–1085.
29. Matsuda H. Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Phys Rev E* 2000;62:3098–3102.
30. Suárez E, Díaz N, Suárez D. Entropic control of the relative stability of triple-helical collagen peptide models. *J Phys Chem B* 2008;112:15248–15255.
31. QSite, version 4.0, Schrödinger, LLC, New York, NY, 2005.
32. Morgunova E, Tuuttila A, Bergmann U, Isupov M, Lindqvist Y, Schneider G, Tryggvason K. Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed. *Science* 1999;284:1667–1670.
33. Becke AD. Exchange-correlation approximation in density-functional theory. In: Yarkony DR, editor. *Modern electronic structure theory part II*. Singapore: World Scientific; 1995.
34. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 2001;105:6474–6487.
35. Murphy RB, Philipp DM, Friesner RA. A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *J Comput Chem* 2000;21:1442–1457.
36. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. *J Phys Chem B* 2001;105:6507–6514.
37. Daura X. Molecular dynamics simulation of peptide folding. *Theor Chem Acta* 2006;116:297–306.
38. Díaz N, Suárez D. Peptide hydrolysis catalyzed by matrix metalloproteinase 2: a computational study. *J Phys Chem B* 2008;112:8412–8424.

Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and QM/MM Calculations

Natalia Díaz,* Dimas Suárez and Ernesto Suárez

diaznatalia@uniovi.es

Supplementary Material

Table S1. Average values and standard deviations for RMSD, RMSF, radius of gyration, and Solvent Accessible Surface Area (SASA) derived from the MD Simulations.

	MMP-2/C9	MMP-2/A13R	MMP-2/C9	MMP-2/A13R
	RMSD (Å)		RMSF (Å)	
All heavy ^a	2.08±0.08	2.09±0.09	0.99±0.10	0.93±0.11
Backbone ^a	1.43±0.09	1.46±0.08	0.66±0.10	0.60±0.11
Helix α2 ^b	0.36±0.04	0.39±0.04	0.23±0.04	0.22±0.04
Helix α3 ^b	0.71±0.06	0.70±0.06	0.26±0.05	0.30±0.07
Strand β4 ^b	0.24±0.04	0.22±0.04	0.13±0.04	0.13±0.04
Zn/Ca S loop ^b	0.95±0.11	1.19±0.13	0.51±0.15	0.45±0.11
β4-β5 loop ^b	0.72±0.27	0.74±0.25	0.46±0.17	0.46±0.17
Ω loop ^b	1.20±0.07	1.19±0.06	0.45±0.08	0.42±0.07
	Radius of gyration^c (Å)		SASA (Å²)	
All heavy	15.08±0.05	15.05±0.04	8936±184	8863±122

^a Without including the *N*-terminal coil (residues Asn₁₁₁-Asn₁₂₂).

^b Corresponding to the backbone heavy atoms.

^c X-ray value (1CK7; catalytic domain only) 14.8 Å.

Table S2. Average values and standard deviations for the RMSF values for the backbone atoms of the **C9** and **A13R** peptide substrates derived from the free- and bound- MD Simulations.

RMSF (Å)	C9	MMP-2/C9	RMSF (Å)	A13R	MMP-2/A13R
Peptide^a	0.52±0.26	1.27±0.37	Peptide^a	1.50±0.49	0.98±0.32
Ser(P₇)	0.16±0.03	0.30±0.03	--	--	--
Gly(P₆)	0.10±0.04	0.51±0.02	Ser(P₆)	0.67±0.02	0.66±0.03
Arg(P₅)	0.07±0.03	0.46±0.02	Gly(P₅)	0.35±0.07	0.41±0.07
Ser(P₄)	0.07±0.03	0.26±0.08	Ala(P₄)	0.18±0.05	0.14±0.06
Leu(P₃)	0.06±0.02	0.06±0.02	Val(P₃)	0.12±0.04	0.07±0.02
Ser(P₂)	0.08±0.03	0.05±0.02	Arg(P₂)	0.12±0.05	0.05±0.02
Arg(P₁)	0.08±0.03	0.04±0.02	Trp(P₁)	0.12±0.05	0.05±0.02
Leu(P₁')	0.08±0.04	0.06±0.03	Leu(P₁')	0.14±0.04	0.06±0.03
Thr(P₂')	0.09±0.04	0.05±0.02	Leu(P₂')	0.15±0.04	0.06±0.03
Ala(P₃')	0.19±0.11	0.24±0.13	Thr(P₃')	0.15±0.05	0.07±0.04
--	--	--	Ala(P₄')	0.25±0.05	0.08±0.04

^a Without including the Ace- and -Nme terminal fragments

Table S3. Geometrical parameters (distances in Å and angles in degrees) characterizing the first shell around Zn₁ along the QM/MM hydrolysis reaction of the **C9** and **A13R** peptide substrates.

		C₁	C₂	TS₁	I₁	TS₂	I₂	TS₃	I₃	TS₄	P
His₄₀₃-Nε···Zn₁	<i>C9</i>	2.12	2.13	2.14	2.12	2.14	2.14	2.12	2.14	2.12	2.09
	<i>A13R</i>	2.15	2.15	2.14	2.14	2.15	2.15	2.15	2.13	2.12	2.11
His₄₀₇-Nε···Zn₁	<i>C9</i>	2.15	2.16	2.18	2.18	2.16	2.16	2.14	2.13	2.13	2.13
	<i>A13R</i>	2.16	2.16	2.19	2.16	2.16	2.15	2.14	2.14	2.14	2.14
His₄₁₃-Nε···Zn₁	<i>C9</i>	2.20	2.25	2.21	2.19	2.15	2.15	2.13	2.16	2.18	2.19
	<i>A13R</i>	2.25	2.31	2.27	2.23	2.20	2.19	2.19	2.18	2.20	2.20
Wat₁-O···Zn₁	<i>C9</i>	2.01	1.99	2.10	2.18	2.12	2.13	2.12	2.14	2.18	2.20
	<i>A13R</i>	2.04	2.00	2.14	2.26	2.18	2.16	2.13	2.15	2.18	2.22
P₁-O···Zn₁	<i>C9</i>	2.34	2.31	2.13	2.06	2.06	2.05	2.09	2.10	2.08	2.11
	<i>A13R</i>	2.28	2.28	2.09	2.01	2.03	2.02	2.05	2.08	2.08	2.07
His₄₀₃-Nε···Zn₁···Nε-His₄₀₇	<i>C9</i>	104	104	100	101	98	99	100	104	106	109
	<i>A13R</i>	99	100	95	96	95	95	97	99	100	101
His₄₀₃-Nε···Zn₁···Nε-His₄₁₃	<i>C9</i>	101	98	100	102	104	103	106	104	104	103
	<i>A13R</i>	104	101	103	105	106	107	107	108	108	107
His₄₀₃-Nε···Zn₁···O-Wat₁	<i>C9</i>	104	102	104	105	107	108	106	110	108	110
	<i>A13R</i>	98	98	100	101	100	99	102	102	101	103
His₄₀₃-Nε···Zn₁···O-P₁	<i>C9</i>	122	123	122	122	120	120	118	120	120	123
	<i>A13R</i>	122	122	122	120	119	118	117	119	120	124
His₄₀₇-Nε···Zn₁···Nε-His₄₁₃	<i>C9</i>	96	93	95	96	98	98	99	98	97	97
	<i>A13R</i>	96	94	95	98	100	101	101	101	100	100
His₄₀₇-Nε···Zn₁···O-Wat₁	<i>C9</i>	91	94	93	92	90	88	88	89	89	89
	<i>A13R</i>	91	93	94	94	92	91	89	90	90	89
His₄₀₇-Nε···Zn₁···O-P₁	<i>C9</i>	134	134	137	135	137	136	136	132	130	126
	<i>A13R</i>	139	138	141	141	141	140	139	136	134	130
His₄₁₃-Nε···Zn₁···O-Wat₁	<i>C9</i>	152	156	153	149	146	146	146	143	144	142
	<i>A13R</i>	155	158	154	149	150	150	148	146	147	146
His₄₁₃-Nε···Zn₁···O-P₁	<i>C9</i>	79	78	84	87	91	91	91	87	88	86
	<i>A13R</i>	80	78	86	88	90	92	91	88	89	88
Wat₁-O···Zn₁···O-P₁	<i>C9</i>	77	81	72	67	63	63	63	62	62	61
	<i>A13R</i>	79	83	72	65	64	63	63	62	62	62

Table S4. Geometrical parameters (distances in Å and angles in degrees) for the reactive events in the QM/MM hydrolysis reaction of the **C9** and **A13R** peptide substrates.

		C₁	C₂	TS₁	I₁	TS₂	I₂	TS₃	I₃	TS₄	P
P₁-C···O-P₁	<i>C9</i>	1.26	1.26	1.29	1.33	1.34	1.34	1.34	1.32	1.31	1.28
	<i>A13R</i>	1.26	1.26	1.30	1.34	1.35	1.35	1.34	1.32	1.31	1.29
P₁-C···N-P₁'	<i>C9</i>	1.33	1.33	1.36	1.41	1.50	1.51	1.56	1.65	1.82	2.44
	<i>A13R</i>	1.33	1.34	1.36	1.42	1.46	1.49	1.57	1.67	1.85	2.45
P₁-C···O-Wat₁	<i>C9</i>	2.47	2.52	1.97	1.64	1.48	1.46	1.44	1.41	1.37	1.28
	<i>A13R</i>	2.61	2.64	1.98	1.60	1.52	1.48	1.43	1.40	1.36	1.28
Wat₁-O···H1-Wat₁	<i>C9</i>	0.98	0.97	0.98	0.99	0.98	0.98	1.00	1.06	1.14	1.75
	<i>A13R</i>	0.98	0.98	0.98	0.99	0.98	0.98	1.00	1.06	1.20	1.74
Wat₁-H1···O-P₂	<i>C9</i>	1.92	2.07	1.93	1.83	2.47	2.72	2.68	2.95	3.06	3.20
	<i>A13R</i>	1.95	2.07	1.99	1.86	2.18	2.51	2.67	2.92	3.02	3.08
Wat₁-H1···Oε1-Glu₄₀₄	<i>C9</i>	2.71	2.66	2.73	2.79	1.90	1.81	1.65	1.50	1.34	1.00
	<i>A13R</i>	2.74	2.69	2.68	2.71	2.22	1.90	1.66	1.47	1.26	1.00
Wat₁-O···H2-Wat₁	<i>C9</i>	1.12	1.38	1.57	1.68	2.39	2.51	2.38	2.27	2.33	2.58
	<i>A13R</i>	1.11	1.38	1.56	1.70	2.12	2.38	2.33	2.28	2.36	2.62
Wat₁-H2···Oε2-Glu₄₀₄	<i>C9</i>	1.30	1.09	1.03	1.01	1.01	1.02	1.17	1.76	1.88	2.10
	<i>A13R</i>	1.34	1.09	1.04	1.01	0.99	1.01	1.25	1.81	1.94	2.22
Wat₁-H2···N-P₁'	<i>C9</i>	3.13	3.26	2.91	2.65	1.93	1.83	1.41	1.05	1.04	1.02
	<i>A13R</i>	3.47	3.54	3.02	2.70	2.29	2.02	1.37	1.05	1.04	1.02
Glu₄₀₄-Oε2···N-P₁'	<i>C9</i>	3.84	3.95	3.61	3.37	2.91	2.84	2.58	2.81	2.92	3.11
	<i>A13R</i>	4.23	4.23	3.74	3.45	3.20	3.00	2.61	2.86	2.97	3.23
P₂-O···O-Wat₁	<i>C9</i>	2.83	2.96	2.86	2.79	2.87	2.88	2.90	2.95	3.01	2.96
	<i>A13R</i>	2.89	3.00	2.92	2.83	2.93	3.01	2.98	3.04	3.09	3.03
P₁-O···P₁-C···Wat₁-O	<i>C9</i>	87	89	98	103	102	102	104	107	110	118
	<i>A13R</i>	84	85	96	103	102	102	104	107	111	118
P₁-C···P₁'-N···Zn₁	<i>C9</i>	50	49	43	42	40	41	39	42	40	36
	<i>A13R</i>	45	44	38	37	37	36	38	38	37	34
P₁-C···P₁'-N···Glu₄₀₄-Cδ	<i>C9</i>	89	89	85	84	92	95	94	92	88	82
	<i>A13R</i>	83	83	82	80	86	90	92	89	85	79
P₁-C···P₁'-N···P₁'-Cα···P₁'-H	<i>C9</i>	164	168	-166	156	-130	-127	-123	-127	-130	-138
	<i>A13R</i>	170	172	-166	156	-137	-132	-122	-127	-129	-139
P₁-C···P₁'-N···Glu₄₀₄-Oε2···Glu₄₀₄-Cδ	<i>C9</i>	46	37	28	25	14	20	11	35	35	45
	<i>A13R</i>	-50	-42	-35	-26	20	16	20	29	27	-37

Table S5. Relative energies (kcal/mol) obtained from the QM/MM Optimizations and Single-Point Calculations on the MMP-2 Catalytic Domain reacting with the **C2** and **A13R** peptides.

	<i>QM/MM</i> ^a		<i>QM</i> ^b		<i>MM</i> ^c		<i>QM/MM</i> ^d		<i>QM/MM</i> + ΔG_{solv} ^d	
	<i>C9</i>	<i>A13R</i>	<i>C9</i>	<i>A13R</i>	<i>C9</i>	<i>A13R</i>	<i>C9</i>	<i>A13R</i>	<i>C9</i>	<i>A13R</i>
<i>C</i> ₁	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>C</i> ₂	-0.71	-0.91	-1.05	-0.81	-1.75	-0.58	-0.49	-1.56	-0.12	-1.13
<i>TS</i> ₁	6.13	10.25	-0.81	6.86	-1.41	-1.53	5.79	7.62	7.58	8.87
<i>I</i> ₁	4.71	6.66	-2.46	6.23	-3.72	-1.78	3.74	3.86	5.53	3.82
<i>TS</i> ₂	11.22	13.03	-0.85	10.24	0.72	-0.36	11.25	10.27	11.58	10.04
<i>I</i> ₂	7.40	12.58	-0.56	10.24	0.65	1.86	5.77	9.47	8.20	8.67
<i>TS</i> ₃	13.75	19.59	-1.34	13.35	3.88	3.21	11.87	16.49	13.23	15.92
<i>I</i> ₃	1.15	6.71	1.86	6.22	1.52	3.83	-2.54	4.34	0.06	3.67
<i>TS</i> ₄	0.04	9.78	1.67	6.59	-0.60	3.12	-5.18	8.15	1.76	8.11
<i>P</i>	-8.20	-1.88	-16.43	-4.59	-0.81	1.95	-12.49	-3.34	-6.16	-4.11

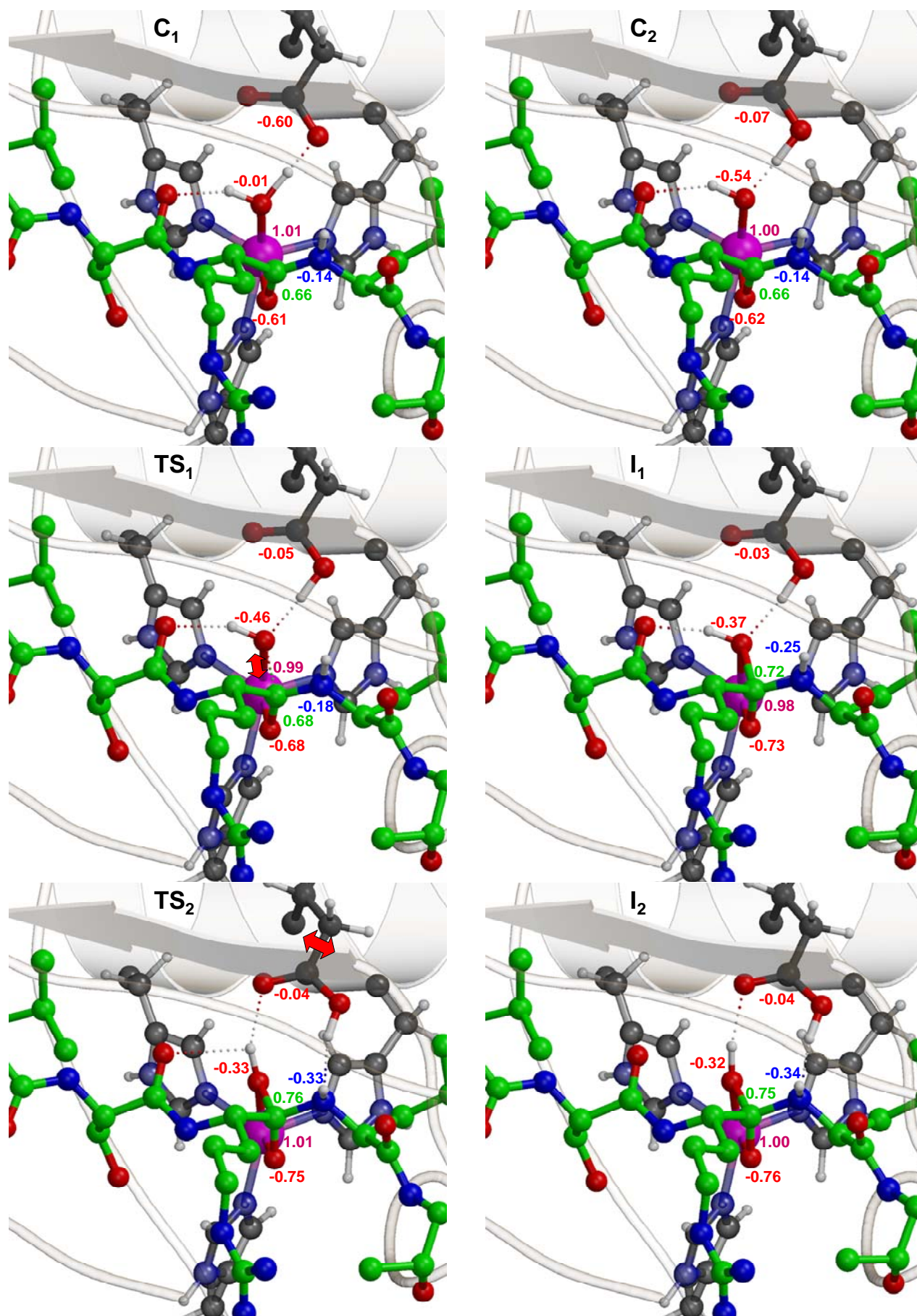
^a Optimized structures

^b From single point calculations on the QM/MM optimize structures including only the QM region

^c MM terms including also the QM-MM interaction

^d From single point calculations including the 50 water molecules closest to Zn₁

Figure S1. Optimized structures for the hydrolysis of C9 in the catalytic domain of the MMP-2 enzyme. Mulliken charges are shown for the reactive atoms including the bound H-atoms.



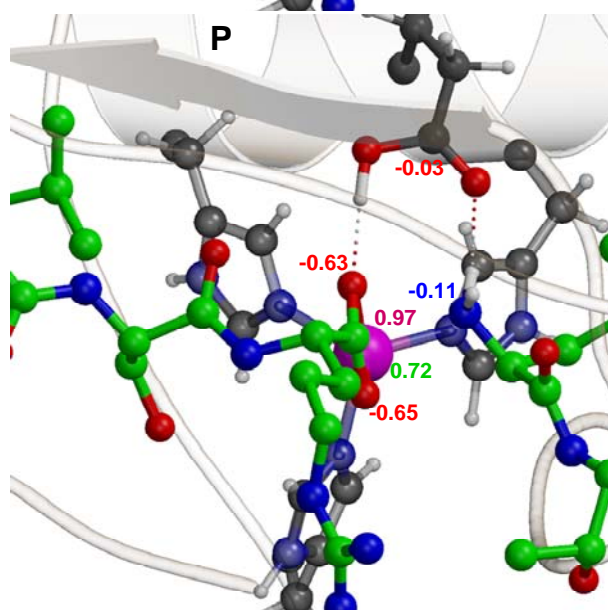
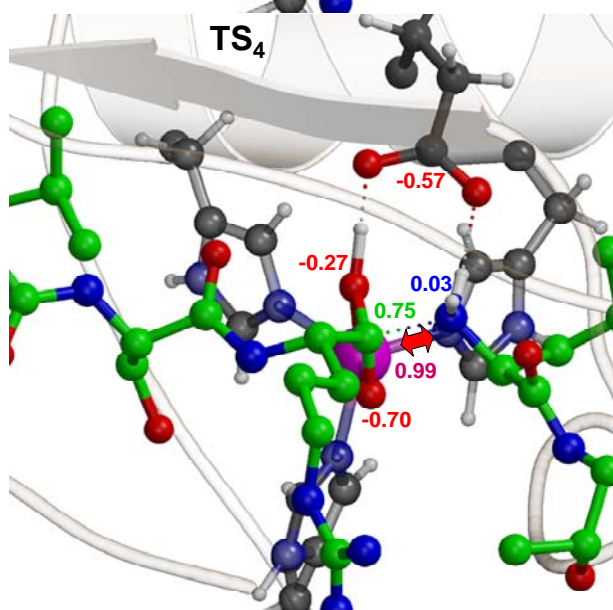
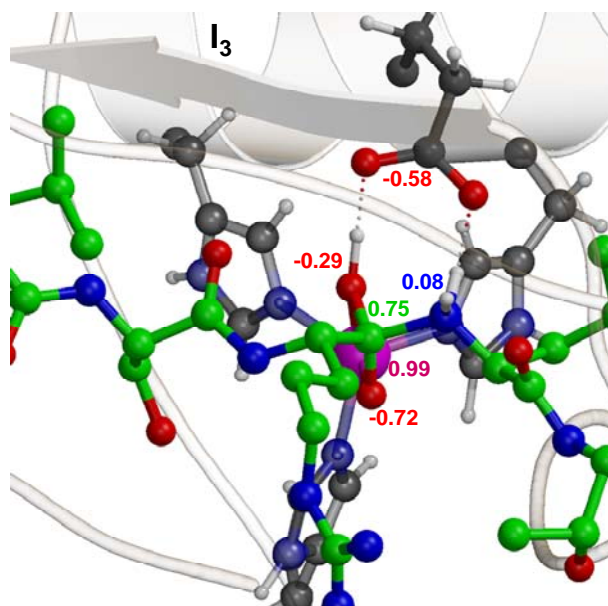
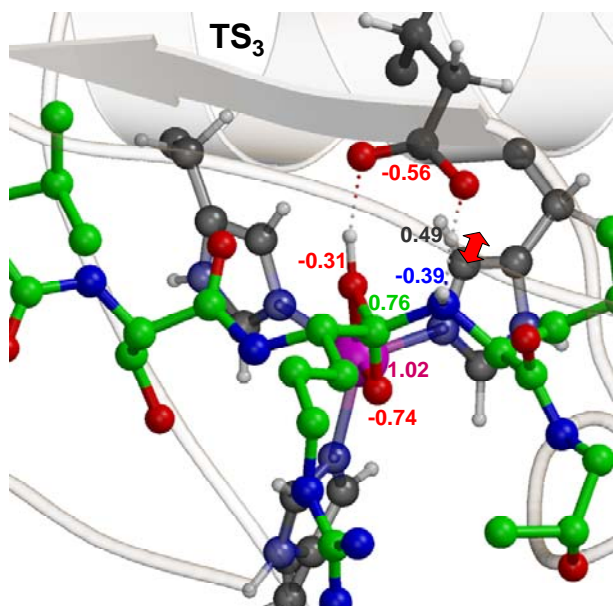
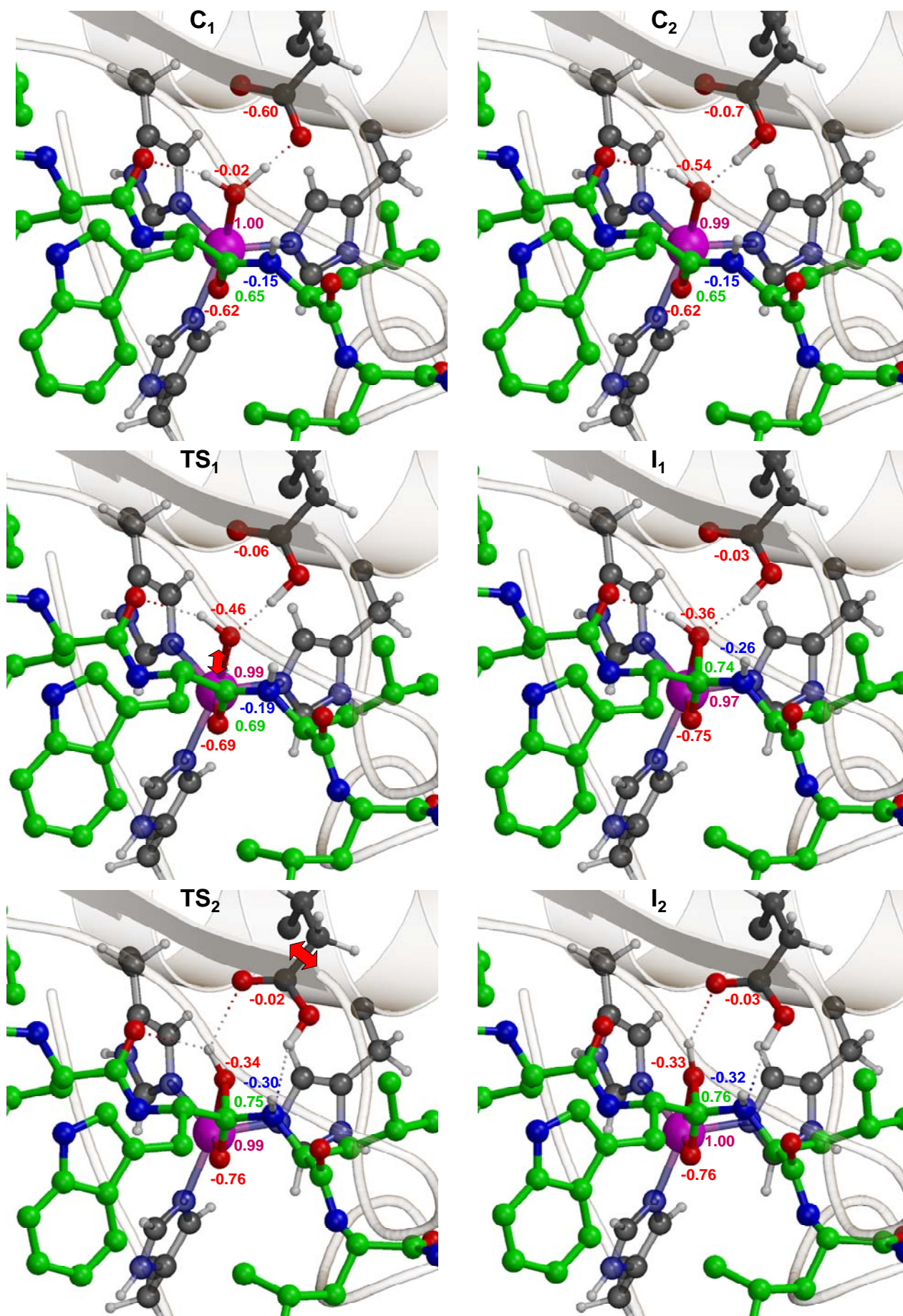
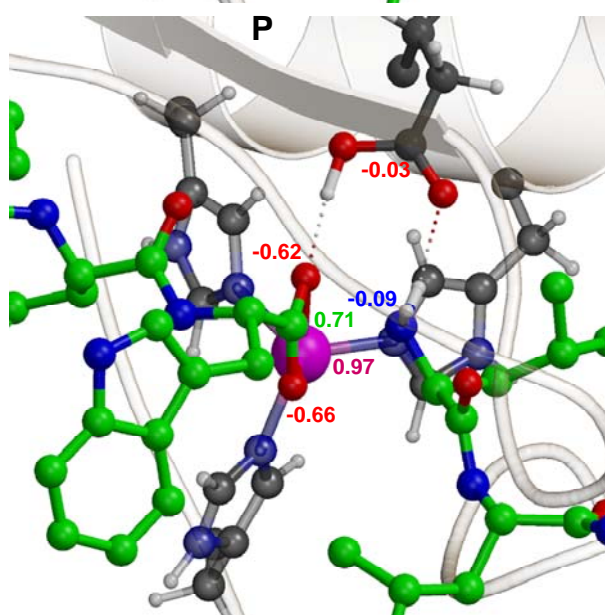
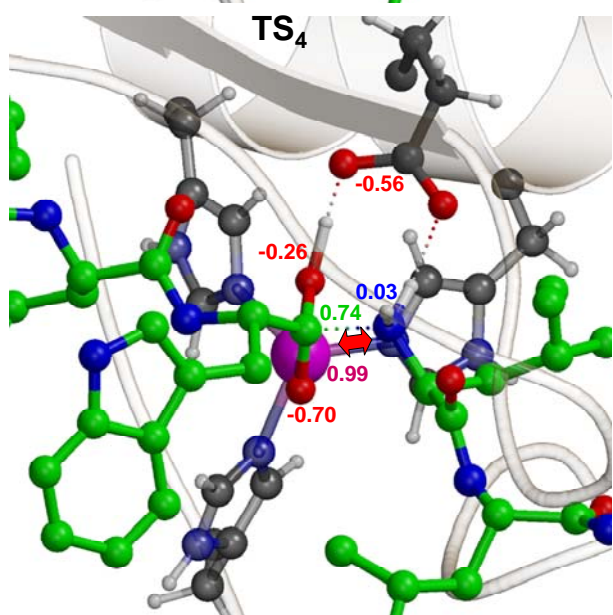
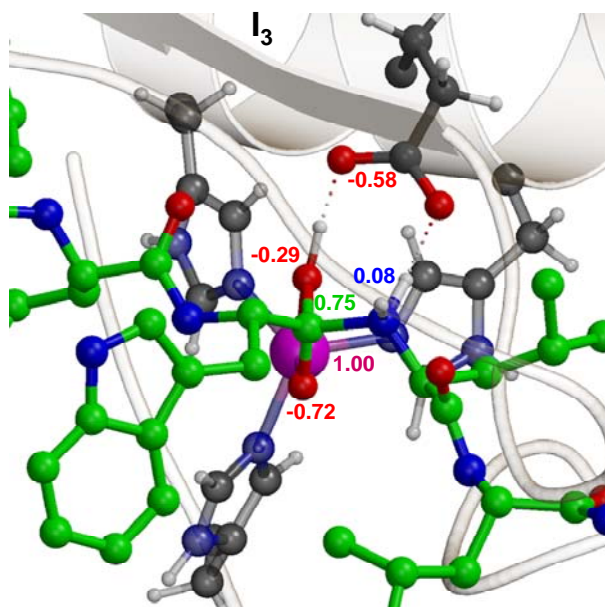
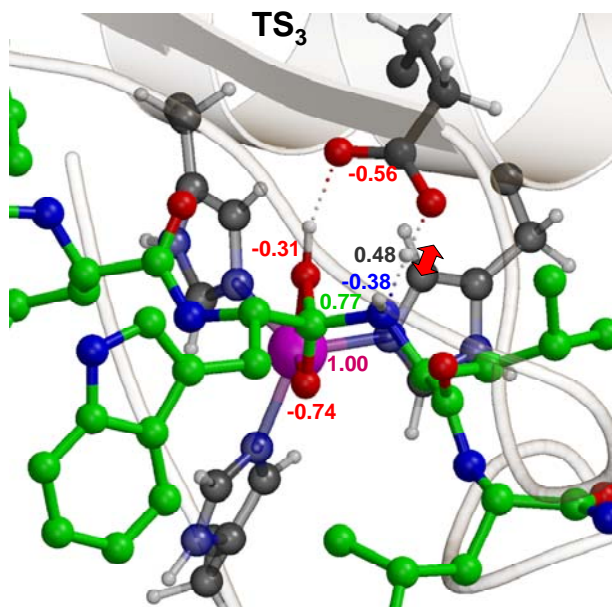


Figure S2. Optimized structures for the hydrolysis of **C13R** in the catalytic domain of the MMP-2 enzyme. Mulliken charges are shown for the reactive atoms including the bound H-atoms.





2.1.1.3 Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations

Ernesto Suárez, Natalia Díaz and Dimas Suárez
J. Chem. Theory Comput. **2011**, 7, 2638-2653

Entropy Calculations of Single Molecules by Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

Julián Clavería 8, Departamento de Química Física y Analítica, Universidad de Oviedo, Oviedo, 33006 Spain

S Supporting Information

ABSTRACT: As shown by previous theoretical and computational work, absolute entropies of small molecules that populate different conformers can be predicted accurately on the basis of the partitioning of the intramolecular entropy into vibrational and conformational contributions. Herein, we further elaborate on this idea and propose a protocol for entropy calculations of single molecules that combines the rigid rotor harmonic oscillator (RRHO) entropies with the direct sampling of the molecular conformational space by means of classical molecular dynamics simulations. In this approach, the conformational states are characterized by discretizing the time evolution of internal rotations about single bonds, and subsequently, the mutual information expansion (MIE) is used to approach the full conformational entropy from the converged probability density functions of the individual torsion angles, pairs of torsions, triads, and so on. This RRHO&MIE protocol could have broad applicability, as suggested by our test calculations on systems ranging from hydrocarbon molecules in the gas phase to a polypeptide molecule in aqueous solution. For the hydrocarbon molecules, the ability of the RRHO&MIE protocol to predict absolute entropies is assessed by carefully comparing theoretical and experimental values in the gas phase. For the rest of the test systems, we analyze the advantages and limitations of the RRHO&MIE approach in order to capture high order correlation effects and yield converged conformational entropies within a reasonable simulation time. Altogether, our results suggest that the RRHO&MIE strategy could be useful for estimating absolute and/or relative entropies of single molecules either in the gas phase or in solution.

INTRODUCTION

A common assumption that lies at the heart of many entropy calculations is that the absolute entropy of a single molecule can be separated into several meaningful contributions.^{1,2} Perhaps the most straightforward and useful division of entropy is into the whole-body translational and rotational ($S_{\text{trans}} + S_{\text{rot}}$) and the intramolecular configurational (S_{config}) contributions, the latter one accounting for the entropy of internal degrees of freedom. Interestingly, the development of reliable and cost-effective strategies for computing the configurational entropy of complex molecular systems is a topic that has received much attention during recent years given that, for instance, the ability to compute accurate ΔS_{config} values could be very useful in understanding both the experimental and theoretical data on folding^{3–8} and/or association^{9–14} of biomolecular systems. Unfortunately, there are still many open questions about S_{config} , concerning its relationship to molecular structure and the importance of correlation among internal motions given that, for relatively large molecular systems, such correlations have been studied only through linear approximations or low-order truncated mutual information expansions.^{15–23}

According to Grubmüller and co-workers,^{24,25} methods that can compute entropy values can be classified into three broad categories: (a) methods based on the computation of free energy differences using thermodynamic integration,²⁶ (b) the hypothetical scanning approach developed by Meirovitch and co-workers,^{27,28}

and (c) an array of direct methods that extract the entropy of a single molecule from configurations generated by carrying out a conventional molecular dynamics (MD) or Monte Carlo simulation. In this work, we are basically interested in the latter category since our approach aims to estimate biomolecular entropies directly from MD simulations.

In what follows, we will briefly review a family of direct methods for entropy calculations that are more relevant to our approach taking into account that these methods can be divided into *parametric* (quasi-harmonic) and *nonparametric* methods depending whether or not they assume a functional form for the probability density function of the internal degrees of freedom. In addition, we also distinguish a third category of direct methods, the hybrid methods that combine elements of the two kinds (i.e., parametric and nonparametric).

Quasi-Harmonic Methods. The original quasiharmonic analysis was the first example of a direct method applied to biomolecular systems. It was first introduced by Karplus and Kushick, showing that the difference in configurational (i.e., non kinetic) entropy between two molecular conformations can be estimated from their respective covariance matrices.¹⁵ The basic idea is to consider the underlying configurational density function $P(\mathbf{q})$ in the classical configurational entropy, $S_{\text{config}} = -k_B \int_C P(\mathbf{q}) \ln$

Received: March 30, 2011

Published: June 30, 2011

$P(\mathbf{q}) d\mathbf{q}$, as a multivariate normal distribution, leading thus to the following entropy expression:

$$S_{\text{config}} = \frac{1}{2} k_{\text{B}} [n + \ln((2\pi)^n \det(\boldsymbol{\sigma}))] \quad (1)$$

where $\det(\boldsymbol{\sigma})$ is the determinant of the covariance matrix of the n internal coordinates. This equation can be applied to estimate only relative entropy values. However, the implementation of the quasi harmonic (QH) method suffers from some practical drawbacks due to the required transformation to internal coordinates and the consequent approximations made in the Jacobian. The need for this transformation comes from the fact that removal of the center of mass translation (unavoidable for convergence reasons) makes $\boldsymbol{\sigma}$ singular if Cartesian coordinates are used.¹⁷

To overcome the limitations of the original QH method, Schlitter has proposed an *ad hoc* approximation to the entropy in which the Cartesian covariance matrix is modified by adding a diagonal matrix so that the resulting $\boldsymbol{\sigma}$ matrix is nonsingular. Starting with the formula for the entropy of a one-dimensional quantum-mechanical harmonic oscillator (HO), Schlitter proposed the following heuristic entropy expression for a system of multiple particles:

$$\begin{aligned} S'_{\text{HO}} &= \frac{1}{2} k_{\text{B}} \sum_i \ln \left(1 + \frac{k_{\text{B}} T e^2}{\hbar^2} \langle q_i^2 \rangle_c \right) \\ &= \frac{1}{2} k_{\text{B}} \ln \left[\prod_i \left(1 + \frac{k_{\text{B}} T e^2}{\hbar^2} \langle q_i^2 \rangle_c \right) \right] \end{aligned} \quad (2)$$

where $\langle q_i^2 \rangle_c$ is the classical variance of the eigenvectors of the mass-weighted covariance matrix $\boldsymbol{\sigma}' = \mathbf{M}\boldsymbol{\sigma}$. Of particular importance is the fact that by adding both translational and rotational entropy contributions to the Schlitter's expression for the configurational entropy, it is possible to estimate absolute rather than relative entropies that are directly comparable to the rigid-rotor harmonic-oscillator (RRHO) entropies obtained from normal-mode analysis and standard statistical thermodynamic formulas.

Following the introduction of Schlitter's method, the QH analysis has been upgraded in order to compute absolute entropies.¹⁸ To this end, the reformulated quasiharmonic approximation (QHA) constructs a pseudoHessian matrix (\mathbf{H}) of the molecular system directly from the Cartesian covariance matrix ($\mathbf{H})_{ij} = k_{\text{B}} T (\boldsymbol{\sigma}^{-1})_{ij}$. The corresponding eigenvectors of \mathbf{H} can be seen to represent motional modes around the average system configuration. By associating each of the quasi-harmonic modes with a one-dimensional harmonic oscillator, the total configurational entropy can be approximated as a sum of harmonic contributions.

Despite the improvements introduced since the original formulation of the QH method, either Schlitter's approach or the renovated QHA method suffer from three potential flaws: (a) only linear correlations are taken into account, and therefore, supralinear correlations among the system variables are ignored; (b) formally, the multim minima potential energy surface is exceedingly smoothed by defining only one minimum and ignoring any anharmonicity (including multimodality) of the essentially multimodal probability density function; (c) there is no clear and unambiguous way to separate the overall rotation from the internal motions (e.g., there are uncertainties up to 80 J/mol K due to the arbitrariness in the choice of reference atoms for the preliminary structure superposition).²⁹ Accordingly, these methods provide an upper limit to the true absolute entropy S_{tot} .³⁰

Furthermore, given a covariance matrix, the function that maximizes entropy is precisely a Gaussian distribution function.³¹

To mitigate their well-known limitations, other authors have proposed several refinements of the QH methods that estimate the importance of the anharmonicity and/or supralinear correlation effects.^{16,19,32–34} For example, three years after Karplus introduced the QH analysis, Berendsen et al. proposed a simple strategy to account for the anharmonicity in the configurational probability density function.^{16,32} In this approach, the configurational part of the classical entropy $S_{\text{config}}(\mathbf{q})$, where $\mathbf{q} = (q_1, q_2, \dots, q_n)$ is an array of internal coordinates, is computed by combining the sum of marginal configurational entropies of the individual q_i variables with the correlation contributions captured by the Karplus model:

$$S_{\text{config}}(q) = -k_{\text{B}} \sum_i \int P(q_i) \ln P(q_i) dq_i + [S_{\text{config}}^{\text{QH}} - S_{\text{config,diag}}^{\text{QH}}] \quad (3)$$

where $S_{\text{config}}^{\text{QH}}$ is the configurational entropy computed by eq 1 and $S_{\text{config,diag}}^{\text{QH}}$ is the uncorrelated or diagonal Gaussian contribution, calculated by zeroing the nondiagonal elements of the covariance matrix.

Very recently, Baron et al. have proposed a more refined method that estimates both the effects of anharmonicity and supralinear correlations on the classical entropy (i.e., not only configurational entropy).^{19,33} The starting point is provided by the renovated QHA approach that uses only Cartesian coordinates.¹⁸ Subsequently, the entropy corrections are estimated from the quasiharmonic coordinates or modes and their corresponding probability densities obtained from the simulations. The classical entropy correction for the nonharmonic behavior, $\Delta S_{\text{cl}}^{\text{ah}} = S_{\text{cl}}^{\text{ah}} - S_{\text{cl}}^{\text{ho}}$, is estimated as the difference between the sum of the marginal entropies obtained directly for the individual modes, $S_{\text{cl}}^{\text{ah}}$, and the entropic term obtained considering the quasiharmonic coordinates as harmonic oscillators, $S_{\text{cl}}^{\text{ho}}$. On the other hand, the pairwise supralinear correction proposed by Baron et al., $\Delta S_{\text{cl}}^{\text{pc}} = \sum_{n>m} (S_{\text{cl,mn}}^{\text{ah}} - S_{\text{cl,m}}^{\text{ah}} - S_{\text{cl,n}}^{\text{ah}})$, is obtained from the classical entropies $S_{\text{cl,mn}}^{\text{ah}}$ computed using the joint probability distribution of the modes m and n , and the corresponding marginal entropies $S_{\text{cl,m}}^{\text{ah}}$ and $S_{\text{cl,n}}^{\text{ah}}$.¹⁹ The estimated entropy finally reads

$$S \approx S_{\text{QHA}} + \Delta S_{\text{cl}}^{\text{ah}} + \Delta S_{\text{cl}}^{\text{pc}} \quad (4)$$

In principle, as remarked upon by the authors, this method can be generalized for the inclusion of higher order correlations.³³ Computational results following this approach have been reported for a microsecond MD trajectory of a peptide model,³³ showing the importance of sufficient phase-space sampling to estimate entropic contributions. It has also been shown that, in accordance with previous studies,^{13,19} the pairwise supralinear correlation is normally large while the effect of the anharmonicity on the entropy calculations is relatively small.

Nonparametric Methods. On the basis of the mathematical tools of information theory and advanced statistics, it is possible to estimate the full-dimensional configurational probability density function $P(\mathbf{q})$ without resorting to any analytical approximation, unlike the QHA and Schlitter methods. This is the case of the method of Hnizdo et al.²⁰ that is based on the use of a series of k th nearest-neighbor (NN) entropy estimators,³⁵ \hat{S}_k , with k being large enough to make a smooth estimation but small enough to make the estimation as local as possible (e.g., $k \in \{1, 2, \dots, S\}$). In

this approach, each molecular configuration is represented as a vector with d components (e.g., the number of internal degrees of freedom). The NN estimation rests on the simple assumption that the configurational probability density function can be approximated *locally* in a nonparametric manner around each sample point q_i using the volume of a d -dimensional sphere centered at q_i and with a radius chosen such that it contains k neighbor data points. Given that the NN \hat{S}_k estimators yield asymptotically unbiased and consistent entropies as the number of data points N increases, they can provide accurate results for any probability distribution provided that sufficiently large samples of molecular simulation data are available. In practice, however, this method has been applied to small molecules due to the large computational cost of the NN searching algorithms when the dimensionality of the problem, d , is larger than 10–15.

Another nonparametric approach has been proposed by Gilson and co-workers based on the mutual information expansion (MIE),^{14,21} which is a systematic expansion of the entropy of a multidimensional system in mutual-information terms of increasing order n that capture the n -body correlations among the molecular internal coordinates. The size of the problem can be reduced up to manageable limits by neglecting all fourth- and higher-order MIE terms, thus allowing the calculation of S_{config} values for several small molecules as well as the change in S_{config} upon binding for protein–ligand systems,^{14,21} but at the cost of sampling millions of molecular configurations for reaching converged results. A combination of the NN and MIE techniques³⁶ has also been proposed.

Hybrid Approaches. Very recently, Hensen et al. have developed a new direct method that combines and improves different techniques with the specific aim of making entropy calculations for relatively large biomolecules feasible.^{24,25} These authors distinguish three parts or blocks in their method. First, they replace the k th NN entropy estimators by adaptive anisotropic ellipsoidal kernels that capture the configurational density in sufficient detail for up to 45-dimensional spaces. Second, they generate minimally coupled subspaces of internal degrees of freedom by applying a linear orthogonal transformation to Cartesian coordinates in such a way that the mutual information among the resulting coordinates is minimized. The new coordinates are subsequently clustered according to their degree of correlation (correlation among different clusters is neglected). Each oversized cluster ($d > 15$) is subdivided into smaller groups with maximum dimensionality $d = 15$, and its configurational entropy is computed as a sum of the estimated entropy of its components (subclusters) and then corrected by means of mutual information functions. For the stiffest degrees of freedom resulting from the orthogonal transformation, Hensen et al. also propose to employ a generalized quasiharmonic Schlitter formula that accounts for their quantum mechanical nature. Thus, the basic idea behind the method of Hensen et al. is that the combination of parametric and sophisticated nonparametric approaches could help overcome many limitations of the QH methods, but without seriously compromising its applicability to relatively large systems.

Clearly, a key element in the adaptive strategy pursued by Hensen et al. is their statistically based clustering of internal degrees of freedom into subsets that are weakly correlated and the separation into *softer* and *stiffer* degrees of freedom. In this respect, other authors have also designed direct methods that assume (*a priori*) a separation between internal degrees of freedom. For example, the method of Thorpe and Ohkubo³⁷ takes advantage of the fact that typical molecular mechanics (MM) Hamiltonians in

implicit solvent are easily separable into *vibrational* (identified as stretching, bending, and improper torsional motions) and *non-vibrational* (identified as torsions) degrees of freedom. These authors show that entropic contributions with respect to an arbitrary temperature for medium-sized systems can be obtained from standard thermodynamical statistical formulas and molecular partition functions which, in turn, can be derived from the density of state functions computed with the weighted histogram analysis method and replica-exchange MD simulations. In this way, it turns out that entropy differences for small- and medium-sized systems can be obtained from conformational and vibrational energy terms. More recently, Brüschweiler and Li²³ have claimed that the configurational entropy is separable into contributions from *hard* and *soft* degrees of freedom (the latter ones identified again with torsion angles) and that correlation effects among the *soft* variables cancel in good approximation on the basis of the results of test calculations on some dipeptide systems. These authors propose then to assess entropy changes between two states of a system at the same temperature, assuming that the entropic contributions of the hard variables do not change. As a first order approximation for conformational entropies is used, this approach is computationally very efficient and can be applicable to protein systems.

Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations.

In previous works, we have investigated the role of conformational entropy in the absolute and relative stability of collagen model peptides²² as well as in the binding of small peptides to the active site of matrix metalloproteases.³⁸ However, the methodological scope of these previous articles was very narrow, as they are focused on the properties of the biomolecular systems being considered. Hence, in this work, we deal with the methodological details of our protocol for entropy calculations of single molecules whose molecular conformational space is sampled by means of classical MD simulations. Following the original proposal by Karplus and co-workers,^{12,39,40} we assume that the total entropy (S_{tot}) of a single molecule (excluding translation and rotation) can be partitioned into a vibrational (\bar{S}_{vib}) and a pure conformational contribution (S_{conform}):

$$S_{\text{tot}} = \bar{S}_{\text{vib}} + S_{\text{conform}} \quad (5)$$

This simple partitioning scheme can be shown to be formally exact as long as one neglects the entropic contributions from the high energy regions between any pair of wells on the potential energy surface of the molecular system.¹² In addition to the entropy partitioning, our approach is further characterized by the two following features:

- First, we employ the harmonic approximation to compute the mean value of S_{vib} over a time series of representative MD snapshots. The required energy minimization and normal-mode calculations can be done even for relatively large molecular systems provided that molecular mechanics or low level quantum mechanical methods are used. Of course, these normal mode and entropy calculations can be carried out within the framework defined by the conventional RRHO approximations and the standard statistical thermodynamic formulas,⁴¹ thus allowing one to estimate absolute entropies.
- Second, the conformational states along the MD trajectory are determined by means of the discretization of the time evolution of internal rotations about single bonds. This transformation, which does not require any *a priori* separation between softer or stiffer degrees of freedom, implies

that the conformational entropy of the whole MD trajectory (S_{conform}) should be naturally computed using the Shannon informational entropy:^{36,42,43}

$$S_{\text{conform}} = k_{\text{B}} \sum_j^{N_{\text{conf}}} (-p_j \ln p_j) \quad (6)$$

where p_j is interpreted as the statistical weight of the j th conformer. However, at this point, we resort to the above-mentioned MIE method in order to approach the full conformational entropy from the converged probability density functions of the individual torsion angles, pairs of torsions, triads, and so on.

The assumption of the entropy partitioning expressed in eq 5 discriminating between vibrational and conformational entropies; the computation of the mean values of the RRHO entropy contributions accounting for the translational, rotational, and vibrational contributions to the absolute entropy; the discretization of the torsional angles; and the concomitant use of MIE for estimating the conformational entropy from data provided by classical MD simulations constitute, altogether, the basic features of the RRHO&MIE entropy method examined in this work. Nevertheless, it must be noticed that some of these methodological ingredients have been employed in previous calculations of the gas-phase entropy of flexible molecules.^{44–49} Thus, it has been shown that the absolute entropy of a *mixture of conformers* can be computed with reasonable accuracy by averaging the RRHO entropy of all of the conformers present at a given temperature and then adding an entropy of *mixing* (ΔS_{mix}) that accounts for the entropic gain in the *mixture* of conformers:

$$S = \bar{S} + \Delta S_{\text{mix}} = \sum_{\alpha} p_{\alpha} S_{\alpha} + R \sum_{\alpha} (-p_{\alpha} \ln p_{\alpha}) \quad (7)$$

where p_{α} , the molar fraction of the α conformer, is typically estimated with the Maxwell–Boltzmann distribution formula in terms of quantum chemical enthalpies or approximate free energies. All of the distinctive conformers including enantiomeric conformers of the same energy are identified using either a direct counting method or automatic conformational search algorithms. Obviously, p_{α} and ΔS_{mix} in the “mixture of conformers” method are equivalent, respectively, to the statistical weight of the α conformer and the conformational entropy S_{conform} in the framework defined by the RRHO&MIE protocol, and consequently, the two approaches would be essentially identical. However, we will see that conceptual and practical differences subsist in the way that the S_{conform} and ΔS_{mix} terms are handled and that the RRHO&MIE approach is more suitable for dealing with relatively large systems.

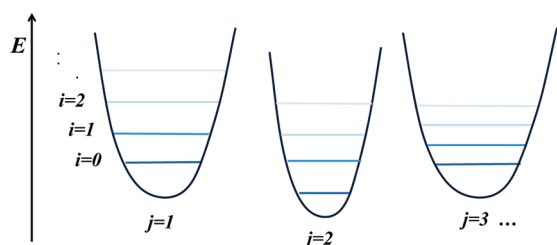
To introduce further details of the RRHO&MIE method and illustrate its potential benefits and limitations, the rest of this paper is organized as follows. First, we will comment on the partitioning of the total entropy (excluding translation and rotation) into the vibrational and conformational components that forms the basis of the RRHO&MIE approach. The separation between vibrational and conformational effects relies heavily on the discretization of the torsional degrees of freedom, and therefore, we will describe this transformation in a detailed manner. Then, after having restated a few definitions about mutual information functions, we will point out that the original MIE expression can be reformulated in such a way that all redundancy in the calculation of the n -order terms is removed. The discretization process combined with the reformulated MIE equation allow us to compute

S_{conform} values including higher order terms beyond the second- or third-order terms that have been considered in most of the previous works. The performance of the RRHO&MIE protocol will be critically discussed on the basis of a series of test calculations on different systems ranging from hydrocarbon molecules in the gas phase to a polypeptide molecule in aqueous solution. For the hydrocarbon molecules (three C_6H_{14} and five C_7H_{16} isomers), their conformational entropies derived from classical MD simulations are combined with their RRHO entropies obtained by carrying out quantum chemical frequency calculations, and the resulting absolute entropies are then compared with experimental data. In this way, we will examine to what extent the assumptions made in the formulation of the RRHO&MIE protocol and the use of classical MD simulations affect the quality of the computed absolute entropies. Second, we will focus on more complex test systems: a series of dipeptide molecules in the gas phase. Although experimental absolute entropies for these test systems are not available, they constitute important cases of study in our validation calculations because they present larger correlation effects among internal degrees of freedom due to the presence of intramolecular H-bond and polar interactions. On the basis of the results obtained for the dipeptide molecules in the gas phase, we will see that the combined discretization process and the MIE approximations are able to capture high order correlation effects and simultaneously yield converged conformational entropies within a reasonable simulation time. Finally, we will analyze the results obtained for a polypeptide molecule in aqueous solution, which can be a representative of the kind of molecular systems for which the RRHO&MIE entropy calculations could be particularly interesting. Besides analyzing the source of conformational correlations, we will also compare the convergence properties and absolute values of S_{conform} as provided by the RRHO&MIE (using an implicit solvent model and MM normal mode calculations) and the QHA methods. Overall, we hope that the methodological proposals and the results of our test calculations could be useful to extend the range of applicability of approximate entropy methods for studying more challenging systems involving polypeptide-folding or molecular association processes.

THEORY AND COMPUTATIONAL METHODS

a. Decomposition of the Intramolecular Entropy. As shown in previous works by Gilson and Zhou,⁴⁰ the entropy of a system with multiple potential energy wells can be formally decomposed into two parts: the entropy that arises from vibrational motions within a single well and the entropy due to conformational transitions between different energy wells. In the case of a single molecule either in the gas phase or in the condensed phase, this entropy decomposition is clearly an approximation whose goodness would depend on the actual molecular structure being studied, temperature, environmental effects, etc. It is true that other entropy methods have been formulated that can derive the total configurational entropy directly from Cartesian coordinates, thus avoiding any assumption about the additivity or lack thereof of the S_{conform} and S_{vib} contributions¹⁹ (although there may remain the problem of separating the overall rotation from the internal motions²⁹). However, although we recognize that the total entropy is strictly a global property, its formal decomposition into the conformational and vibrational terms constitutes the basis for the present work. As a matter of fact, many experimental and theoretical methods have provided meaningful results by accepting that free

Scheme 1. Schematic Example of a Multiminima Potential Energy Surface



energies or entropy contributions can be attributed to particular degrees of freedom and/or physical interactions.^{1,2}

In the Supporting Information, we provide an alternative derivation of the formal decomposition of the single molecule entropy assuming that the translational and rotational degrees of freedom have been removed and that the resulting potential energy surface in terms of the remaining internal degrees of freedom can be approximated by a set of *distinguishable* energy basins, as shown in Scheme 1. The final entropy decomposition formula is

$$S_{\text{tot}} = \sum_j p_j S_{\text{vib}}^j + S_{\text{conform}} \quad (8)$$

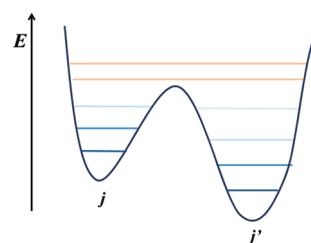
where p_j and S_{vib}^j stand for the probability and vibrational entropy associated with the j th energy basin, respectively, while S_{conform} is the conformational entropy that arises from the populations of the different energy basins.

The practical implementation of the entropy decomposition (eq 8) implies that each molecular configuration of a single molecule (e.g., one MD snapshot) employed in the entropy calculations should be associated with one molecular j conformer (or j energy basin) representing its conformational state. In our protocol, this assignation is only relevant for the evaluation of the conformational entropy S_{conform} and can be achieved by discretizing the time evolution of the torsion angles (see below). The average value of S_{vib} is obtained separately by means of energy minimization calculations followed by normal-mode analyses within the context of the harmonic oscillator model. In terms of accuracy, the formal entropy decomposition should provide nearly exact results for small molecules in the gas phase (ideal conditions) at room temperature given that, in this case, only the lowest vibrational levels would be populated and they could be tagged univocally to a single conformer.

Could the entropy decomposition provide meaningful results in other more complicated situations such as that depicted in Scheme 2? As temperature increases, high energy vibrational levels that can be simultaneously assigned to different conformers would become populated, partially blurring the distinction between vibrational and conformational motions. Although the practical implementation of eq 8 would not be impeded in this case, the entropy given by eq 8 would be overestimated due to the double-counting of the contributions of those high energy vibrational levels accessible from the two energy basins. Therefore, we propose that, in general, the decomposition of the single molecule entropy could be useful, as it would provide an upper bound to the actual value. Nonetheless, its performance needs to be assessed by carrying out test calculations.

b. On the Use of the Harmonic Oscillator Model for Vibrational Entropies. We think that one remarkable advantage of discriminating between vibrational and conformational entropies is the utilization of the HO model for computing the average

Scheme 2. High Energy Vibrational Levels Can Be Assigned to Different Energy Basins



vibrational contribution over a series of representative structures. Clearly, a straightforward computational protocol can be applied for obtaining the mean values of S_{vib} . Starting with a set of representative MD snapshots, optimization calculations relax the internal geometry of each molecular structure to that corresponding to a particular j energy basin, whose vibrational entropy, S_{vib}^j , is subsequently estimated by means of conventional normal mode calculations. In fact, the combination of the HO approximation, the rigid rotor model, and the standard formulas of statistical thermodynamics based upon canonical partition functions can be used to estimate absolute entropies that in some cases admit a direct comparison with experimental data (e.g., in the gas phase).⁵⁰ We can also benefit from many efficient implementations for performing either geometry optimizations and second derivative calculations using molecular mechanics⁵¹ and/or quantum mechanical methods depending on the size of the molecular system. Part of the limitations of the HO model (e.g., the lack of anharmonic effects, errors arising from the level of theory, etc.) could be mitigated by using empirical corrections in the form of scaling factors.^{52,53} We also note that the RRHO entropies (complemented with the conformational contributions) could be useful within the context of approximate free energy methods like the so-called molecular mechanics Poisson–Boltzmann method.⁵⁴ Moreover, previous computational experience with these approaches has shown that the average normal mode entropy converges quite well in terms of the length of simulations and the number of molecular configurations required.^{54,55}

c. Discretization of the Torsional Degrees of Freedom. To take into account the S_{conform} contribution, which arises from the population distribution of the molecular conformers, we propose to discretize the probability density functions of those torsion angles that are commonly used to define the molecular conformational state. To this end, we apply the following protocol.

First, we collect the time series with the values of the torsion angles along the MD simulation. For each torsion angle θ , we have a sample of size N , $\{\theta_1, \dots, \theta_N \mid \theta_i \in [0, 2\pi)\}$, where N is the number of MD snapshots ($\sim 10^5$ – 10^6). To obtain an analytic representation of the underlying probability density function for the torsion θ , we employ the von Mises kernel estimator (the von Mises density is the circular analogue of the Gaussian distribution). Specifically, the probability density function of a given torsion, $\rho(\theta)$, is then approximated by the arithmetic mean of N von Mises distributions centered on the θ_i values:

$$\hat{\rho}(\theta; v) = \frac{1}{2\pi N I_0(v)} \sum_{i=1}^N \exp\{v \cos(\theta - \theta_i)\}$$

where $I_r(v)$ is the modified Bessel function of order r , and v , which is the so-called concentration parameter, is the inverse of the smoothing parameter of the kernel estimator. The value of v

is obtained by applying the recently derived “von Mises-scale plug-in rule”³⁶ for the smoothing parameter, which results in the following expression that depends on the number N of data points:

$$v = [3N\hat{\kappa}^2 I_2(2\hat{\kappa}) \{4\pi^{1/2} I_0(\hat{\kappa})^2\}^{-1}]^{2/5}$$

with $\hat{\kappa}$ being an estimate of the concentration parameter of the global data, for which we take $\hat{\kappa} = 1$, as this value leads to a slightly oversmooth distribution that is more convenient for our purposes. By setting an empirical value for $\hat{\kappa}$ rather than for v , we keep the dependence on N .

The advantage of using the von Mises kernel estimator instead of a normalized histogram method is that we can characterize the analytical properties of $\rho(\theta) \approx \hat{\rho}(\theta;v)$ in order to automatically optimize the location of the maximum and minimum values of $\rho(\theta)$, which is a prerequisite for discretizing the time evolution of the torsion angle θ . This task is performed by the analytical evaluation of the first and second derivatives of $\hat{\rho}(\theta;v)$ over a grid of $\phi_k = k(\pi/(180))$ points with $k = 0, 1, \dots, 359$. On one hand, the approximate positions of the $\hat{\rho}(\theta;v)$ critical points are first determined by averaging two consecutive grid points k and $k + 1$ for which $\hat{\rho}'(\phi_k;v) \hat{\rho}'(\phi_{k+1};v) \leq 0$. For the torsional distributions, the maximum critical points are easily identified thanks to their largely negative second derivative, and therefore, we first locate the maxima of $\hat{\rho}(\theta;v)$ and use the intermediate position between two consecutive maxima as the initial guess for searching the minima. The minima of $\hat{\rho}(\theta;v)$ are found by means of a steepest descent search that adopts a convergence threshold for the residual gradient of 10^{-4} and employs a linear interpolation approximation for evaluating the gradient on the basis of the $\hat{\rho}'(\phi_k;v)$ values.

Once the $\theta_{\min,i}$ values corresponding to the $\hat{\rho}(\theta;v)$ minima are found (let us suppose that there are m minima and $\theta_{\min,i} < \dots < \theta_{\min,m}$), the configurational space of θ defined by the $[0, 2\pi)$ interval is divided into m nonoverlapping intervals ($[\theta_{\min,1}, \theta_{\min,2})$, \dots , $[\theta_{\min,m-1}, \theta_{\min,m})$, $[\theta_{\min,m}, 2\pi) \cup [0, \theta_{\min,1})$) that, in turn, define the different conformational states accessible to θ . In this way, the initial time series containing N data points, $\{\theta_1, \dots, \theta_N\}$, is easily transformed into a set of N integer numbers $\{a_1, \dots, a_N\}$ labeling the conformational states populated by the torsion angle. For example, if θ corresponds to an internal rotation about a C(sp³)–C(sp³) bond, its associated a_i variable could have values of 1, 2, and 3 representing the $g+$, $g-$, and $anti$ conformations, respectively. Therefore, the continuous variable θ characteristic of the torsion angle becomes a discrete random variable A , whose probability mass function, $P(A)$, can be estimated by the maximum likelihood method fed with its corresponding outcomes $\{a_1, \dots, a_n\}$. Finally, we note that the loss of entropy during the $\theta \rightarrow A$ transformation can be expected to be *vibrational*, and that we assume that such a contribution can be reasonably accounted for by normal mode calculations. In other words, the entropy due to the fluctuations of the torsion angles around a local minimum should be recovered by the S_{vib} calculations. This means that, in our approach, it is not necessary to distinguish between *soft* or *hard* degrees of freedom (or between *vibrational* or *nonvibrational* ones) because the conformational entropy turns out to be purely *informational* as a consequence of the discretization process. The probability density function for the torsion angle θ about the C2–C3 bond of 2-methyl-hexane is shown in Figure 1.

d. Mutual Information Expansion: Application to Conformational Entropy Calculations and Reformulation into a Computationally Efficient Scheme More Suitable for Very Large Systems. As shown in the previous section, the conforma-

tional state of a torsion angle can be associated with a one-dimensional random variable A . Analogously, the conformational state of a set of M torsion angles can be described by an M -dimensional random vector (A_1, \dots, A_M) , or alternatively, the conformational state can also be associated with an ordered set $\{A_1, \dots, A_M\}$, where A_i specifies the conformational state of the i th torsion and M is the size in terms of the number of torsion angles of our system $\mathcal{A} = \{A_1, \dots, A_M\}$. Of course, for medium-sized and large molecules, the number of potentially accessible conformers is huge ($\sim 3^M$), and in this case, obtaining the underlying probability mass function $P(A)$ is practically impossible due to sampling limitations. As other authors have done previously,^{21,36} we will use the mutual information expansion as a workaround to this problem. The basic idea here is that if we are able to obtain converged values of the probability mass functions of the individual torsion angles $p(A_i)$, pairs of torsions $p(A_i, A_j)$, triads $p(A_i, A_j, A_k)$, and so on, then we can approach the full-dimensional informational entropy of the (A_1, A_2, \dots, A_M) variables (i.e., the conformational entropy) by including systematically n -order correlations among the A_i variables as measured by mutual information functions.⁴² More specifically, the mutual information expansion leading to the total entropy can be written as

$$S(A_1, \dots, A_M) = \sum_{i=1}^M S(A_i) - \sum_{i < j} I_2(A_i, A_j) + \sum_{i < j < k} I_3(A_i, A_j, A_k) - \sum_{i < j < k < l} I_4(A_i, A_j, A_k, A_l) + \dots$$

where $S(A_i)$ is the informational entropy of the i th torsion angle along the MD simulation while $I_2(A_i, A_j)$, $I_3(A_i, A_j, A_k)$, and so forth are the corresponding mutual information functions that capture the general dependence among the (A_1, A_2, \dots, A_M) variables (unlike the covariance function used in the QH methods that only measures linear correlations). More specifically, the mutual information shared by two variables, A_i and A_j , is computed by combining the informational entropies of the single- and two-variable probability mass functions of the torsion angles:

$$I_2(A_i, A_j) = S(A_i) + S(A_j) - S(A_i, A_j)$$

where $S(A_i, A_j)$ is the joint entropy of A_i and A_j . Both $S(A_i)$ and $S(A_i, A_j)$ can be computed using a Shannon-type expression, $-k_B \sum p_\alpha \ln p_\alpha$ where p_α is the corresponding probability mass function and the sum runs over the possible states accessible to the individual torsion angles A_i or to the pair of torsions (A_i, A_j) . Similarly, the third-order function $I_3(A_1, A_2, A_3)$ includes correlation effects among three torsion angles:

$$I_3(A_i, A_j, A_k) = S(A_i) + S(A_j) + S(A_k) - S(A_i, A_j) - S(A_i, A_k) - S(A_j, A_k) + S(A_i, A_j, A_k)$$

In general, the mutual information shared among k variables $\mathcal{F} = \{A_1, A_2, \dots, A_k\}$ can be generalized by

$$I_k(\mathcal{F}) = \sum_{l=1}^k (-1)^{l+1} \sum_{\substack{\mathcal{F} \subset \mathcal{F} \\ |\mathcal{F}|=l}} S(\mathcal{F})$$

where for every value of l , the inner sum runs over all possible $\binom{k}{l}$ subsets of \mathcal{F} with l elements, $|\mathcal{F}|$ being the cardinality of \mathcal{F}

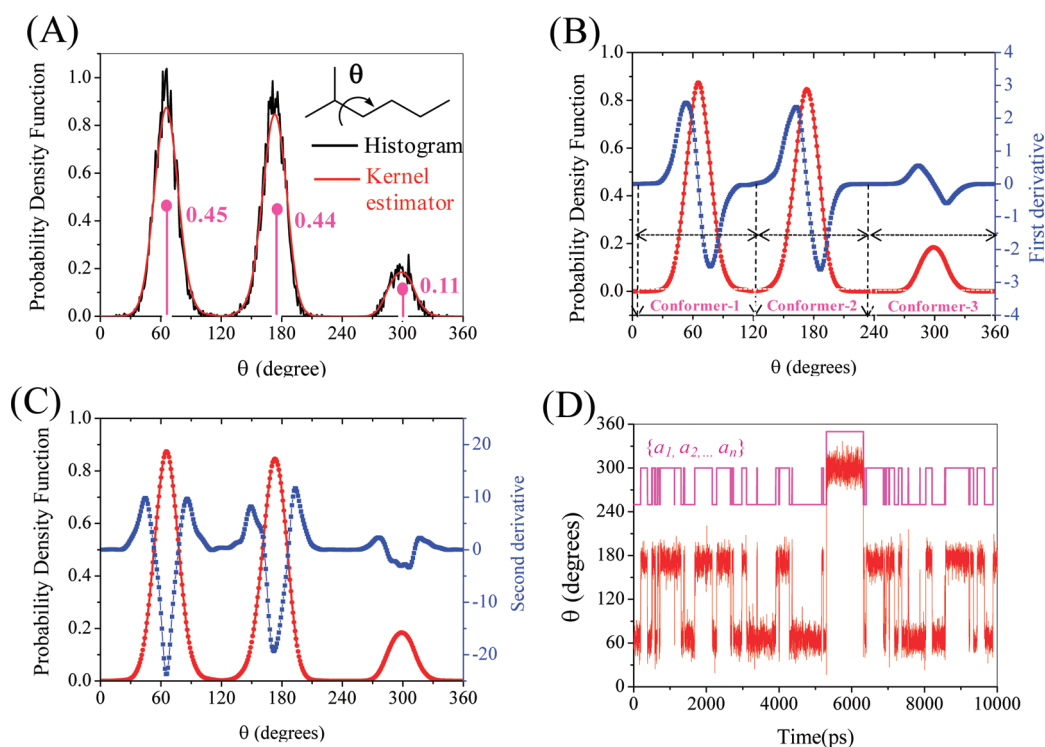


Figure 1. (A) Probability density function for the torsion angle θ about the C2–C3 bond of 2-methyl-hexane as obtained from a histogram representation and a Von Mises kernel estimator. The probability mass function of the three conformational states is also indicated by the vertical bars. (B and C) Superposition of the probability density function and its first and second derivatives as estimated by the Von Mises kernel. (D) Time evolution of the torsion angle θ and its associated discrete variable A (see text for details).

(i.e., the number of elements). A special case arises when $k = 1$, that is, $I_1(J_1) = S(J_1)$. The total entropy of the M variables $\mathcal{A} = \{A_1, \dots, A_M\}$ is usually approximated by a truncated mutual information expansion up to order n ,^{21,42} which becomes an exact expression when $n = M$:

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}) \quad (9)$$

However, we note again that in previous calculations of configurational entropies based on the MIE approach, only low order (i.e., $n \leq 3$) approximations have been used.

A computational shortcoming of the usual form of the MIE is that, for any subset \mathcal{J} of \mathcal{A} with cardinality lower than n , its entropy must be evaluated more than once. If the order of the expansion is not too high and a reasonable amount of rapid access memory is available, all of the required $S(\mathcal{J})$ terms can be stored and used repeatedly in the calculation of the high order terms. For large and strongly correlated systems (e.g., $M > 100$, $n = 3-5$), however, this simple approach could become unfeasible due to memory depletion, while a direct implementation in which the $S(\mathcal{J})$ terms would be recomputed on-the-fly as needed would be prohibitively expensive. In order to remove all redundancy in the calculation of the MIE terms, the original expression can be reformulated in the following manner:

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=k}} S(\mathcal{J}) \quad (10)$$

where the entropy of each subset \mathcal{J} is computed and used only once. The formal proof of the equivalence between eqs 10 and 9 is given in the Supporting Information. To the best of our knowledge, the above expression has not been reported so far, although its implementation could greatly simplify the MIE computational problem for very large systems with hundreds or even more torsion angles. Finally, we note that all of the MIE calculations reported in this work were carried out using a FORTRAN90 code that has been developed in our laboratory and that will be reported elsewhere.⁵⁷

e. Molecular Dynamics Simulation Settings. *Small Systems: Gas-Phase MD Simulations.* MD calculations for each individual molecule were carried out by means of the Amber10 package.⁵⁸ The generalized AMBER force field⁵⁹ was used for the alkane molecules, while the dipeptides (Ace–X–X–Nme, with X = Ala, Ser, Asn, Leu, and Lys) were represented by the AMBER03 force field.⁶⁰ To derive the atomic charges of the alkane molecules, we performed HF/6-31G(d) geometry optimizations of the fully extended conformers followed by single-point B3LYP/cc-pVTZ calculations using the Gaussian 03 program.⁶¹ Other parameters were generated automatically using the antechamber module included in the Amber10 package. Subsequently, 2.0 μ s MD trajectories for all of the alkane and dipeptide compounds were run in the gas phase at 298 K and 1.0 atm using a 1.0 fs time step. Coordinates were saved every picosecond of simulation time (2×10^6 structures).

Polypeptide System: MD Simulation in Solution. We simulated the following hexapeptide sequence, Ace–Pro–Phe–Glu–Leu–Arg–Ala–NH₂ (termed the PFG peptide), which corresponds to one of the peptide sequences that has been selected from a peptide library mixture for probing the cleavage

site motifs of matrix metalloproteinases (MMPs).⁶² Starting coordinates were obtained from conformational search calculations using the LMOD program⁶³ linked to the Amber package. In the LMOD calculations, we employed the AMBER03 force field coupled with the Hawkins–Cramer–Truhlar pairwise generalized-Born (HCT-GB) model.⁶⁴ The lowest energy LMOD structure of PFG was then surrounded by a periodic truncated octahedral box of TIP3P water molecules that extended ~ 12 Å from the protein atoms (~ 1400 water molecules). The solvent molecules were initially relaxed by means of energy minimizations and 50 ps of MD. Subsequently, the full system was minimized and heated gradually to 300 K during 50 ps of MD. During the MD simulation, the system remains coupled to a thermal and a hydrostatic bath at $T = 300$ K and $P = 1.0$ atm, the time step of integration was 2.0, the SHAKE procedure on the X–H bonds was applied, and the particle-mesh-Ewald approach was used for nonbonded interactions. A 2.0 μ s trajectory was computed, and coordinates were saved for analysis every picosecond.

f. Normal Mode Calculations. *Alkane Molecules: Quantum Mechanical Calculations.* For the alkane molecules, we evaluated the average value of their rotational and vibrational entropies by carrying out quantum mechanical (QM) calculations on selected MD snapshots. To this end, 2000 equally spaced snapshots were extracted from the MD trajectories of each compound. All of the MD snapshots were minimized and scored in terms of their relative MM energies. The relative energies of the relaxed structures were compared in order to filter out all of the energetically equivalent structures, obtaining thus a relatively small set of energetically distinguishable molecular conformers (i.e., this means that only one structure for each pair of enantiomeric conformers is retained). The statistical weight of each energetically unique conformer was estimated by its relative abundance in the initial data set containing the 2000 structures. These conformers were subsequently minimized at the B3LYP/cc-pVTZ level of theory^{65,66} and further characterized by analytical frequency calculations. Then, thermal contributions to the gas-phase entropy of the translational, rotational, and vibrational degrees of freedom were obtained within the context of the RRHO approximation and using the B3LYP/cc-pVTZ moments of inertia and vibrational frequencies. Entropy contributions of overall rigid-body rotation take into account the corresponding external symmetry number for each conformer.

Polypeptide Systems: Normal Mode MM Calculations. Taking into account the relatively large size of the polypeptide system, we decided to use MM methodologies for carrying out the required normal mode calculations. Moreover, since the solute and solvent fluctuations are coupled to each other during the MD simulation of the TIP3P water box, we also used the HCT-GB implicit solvent model for removing the explicit consideration of solvent degrees of freedom. Thus, we extracted 10 000 equally spaced snapshots from the 2.0 μ s trajectories of PFG. These structures were postprocessed through the removal of all solvent and counterion molecules. Then, solute entropic contributions were estimated for each structure using the NAB package.⁶⁷ Prior to the normal mode calculations, the geometries of the system described by the AMBER03 force field were minimized until the root-mean-squared deviation of the elements in the gradient vector was less than 10^{-5} kcal/(mol Å). It may be interesting to note that the large majority of the minimized structures (>98%) corresponded to different minima on the potential energy surface, as expected from the relatively large size of this system. These minimizations and the subsequent normal mode calculations⁵¹ were carried out

using the HCT-GB solvent model. Finally, the RRHO entropic contributions were averaged over the 10 000 snapshots.

RESULTS AND DISCUSSION


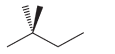
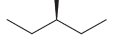
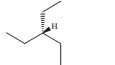
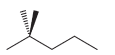
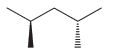
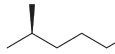
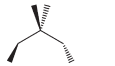
Absolute Entropies of Alkanes. Table 1 and Figures 2 and 3 collect the results of our entropy calculations for the C_6H_{14} (1–3) and C_7H_{16} (4–8) isomers. Of particular interest in these test calculations is the fact that the experimentally available gas-phase entropies of the alkane molecules can be directly compared with our theoretical estimates, since we employ both the RRHO approximation and standard statistical formulas for obtaining the translational–rotational–vibrational entropies of polyatomic molecules,⁴¹ which are subsequently complemented by adding the conformational entropy contributions arising from the discretization of the time evolution of the torsion angles during the classical MD simulations. Furthermore, consideration of these molecules, which are still small enough for exhaustively exploring their conformational space, allows us to show more clearly the relationship of our method with the closely related protocol of the “mixture of conformers” model, which has provided accurate entropy values of small molecules.^{44,45}

Automatic checks of the relaxed snapshots show that the number of energetically distinguishable conformers (N_E in Table 1) that are populated at 298 K in the gas phase vary between 4 and 15 for the various alkane molecules, except for 2,2-dimethyl-butane (2), which shows a single energy level because torsional motions of the methyl groups and about the C2–C3 bond interconnect structures that are degenerate in terms of their potential energy. Precisely, the RRHO entropy value of 2,2-dimethyl-butane at the B3LYP/cc-pVTZ level, 357.77 J/(K mol), is only less than 1 J/(K mol) below the experimental value (378.65), thus showing that the B3LYP/cc-pVTZ level of theory accounts well for the majority of the gas-phase entropy of alkane molecules having a single conformer. This also suggests that frequency scaling, which has been commonly used to correct for systematic errors in the ab initio computation of vibrational frequencies⁵² is probably not required at the B3LYP/cc-pVTZ level. For the rest of the test compounds, however, several energy-distinguishable conformers are significantly populated at 298 K. Their individual RRHO entropy values are similar, but they exhibit non-negligible differences as large as 5 J/(K mol).

The mean values of the RRHO absolute entropies (\bar{S} in Table 1) underestimate the experimental data up to tenths of a J/(K mol) for the flexible hydrocarbon molecules. The correlation plot shown in Figure 2A shows more clearly the differences between the experimental and the average theoretical entropies: the resulting squared correlation coefficient is quite low, $R^2 = 0.70$ (a unit slope is imposed), and a large offset at the intercept arises (-14 J/(K mol)). Moreover, the relative entropy values among the C_6H_{14} or C_7H_{16} isomers are poorly described by the average RRHO values: for example, the ΔS_{exp} value between 1 and 2 is -30.17 J/(K mol), whereas the computed $\Delta \bar{S}$ value is only -7.89 J/(K mol). Therefore, it is clear that neglecting conformational entropy contributions significantly affects the quality of the entropy calculations even for simple hydrocarbon molecules.

Concerning the conformational entropies of the alkane molecules estimated from the classical MD simulations, we first assess their statistical convergence by plotting the S_{conform} values vs simulation time at various expansion orders for 2-methyl-hexane (7) and 3,3-dimethyl-pentane (8; see Figure 3). Thus, all of the S_{conform} profiles in Figure 3 converge to nearly zero-slope curves

Table 1. Average Value of the Translation-Rotational-Vibrational Entropy (\bar{S}) and Converged Values of the Conformational Entropy at 298 K (values in J/mol K) for the Alkane Compounds Studied in This Work^a

Alkane molecules	N_E	S_{exp}^b	\bar{S}^c	$S_{conform}^d$	Ω_{term}	$S_{conform}^{term}$	$\bar{S} + \Delta S_{conform}^e$
1 hexane 	9	388.82 ±0.84	365.94	32.44	3 ²	18.27	380.11
2 2,2-dimethyl-butane 	1	358.65 ±0.84	357.77	45.67	3 ⁵	45.67	357.77
3 3-methyl-pentane 	5	382.88 ±0.67	365.27	43.11	3 ³	27.40	380.98
4 3-ethyl-pentane 	8	411.50	393.78	48.7	3 ³	27.40	415.08
5 2,2-dimethyl-pentane 	4	392.88	388.56	46.13	3 ⁵	45.67	389.02
6 2,4-dimethyl-pentane 	4	396.73	386.38	43.56	3 ⁴	36.54	393.40
7 2-methyl-hexane 	15	419.99	397.25	45.12	3 ³	27.40	414.96
8 3,3-dimethyl-pentane 	4	398.02	381.90	52.35	3 ⁴	36.54	397.71

^a The number of energetically indistinguishable conformers at $T = 0$ (Ω_{term}) and the number of energetically distinguishable conformers (N_E) at 298 K are also indicated. ^b From references: hexane,⁶⁸ 2,2-dimethyl-butane,⁶⁹ 3-methyl-pentane,⁷⁰ 3-ethyl-pentane,⁷¹ 2,2-dimethyl-pentane,⁷¹ 2,4-dimethyl-heptane,⁷¹ 2-methyl-hexane,⁷¹ and 3,3-dimethyl-pentane.⁷⁰ ^c $\bar{S} = S_{trans} + S_{rot}^{RR} + S_{vib}^{HO}$ and using the B3LYP/cc-pVTZ geometries and frequencies. The S_{rot} and S_{vib} contributions are averaged according to the relative frequency of the energetically distinguishable conformers during the classical AMBER MD simulations. ^d Including the fifth-order MIE conformational entropy derived from the classical MD trajectories. ^e $\Delta S_{conform} = S_{conform} - S_{conform}^{term}$ (see text for details).

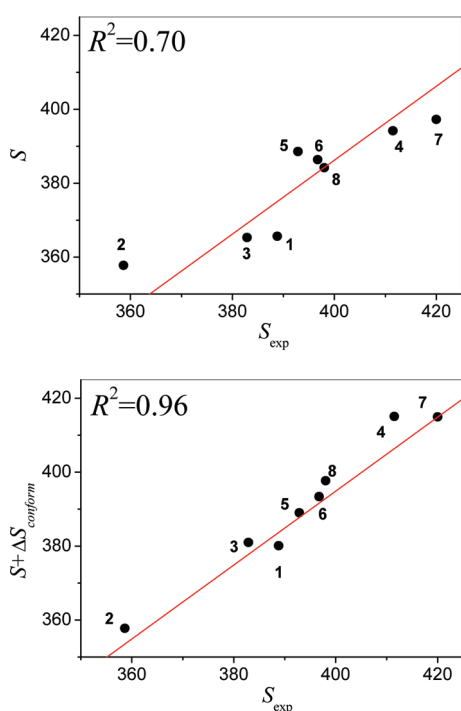


Figure 2. Correlation plots between theoretical absolute entropies (in J/(K mol)) for the eight alkane molecules considered in this work before (A) and after (B) adding the conformational entropy contribution to the mean RRHO entropies.

with respect to simulation time after ~ 1.5 and ~ 1.0 μ s for 7 and 8, respectively, regardless of the expansion order. An even

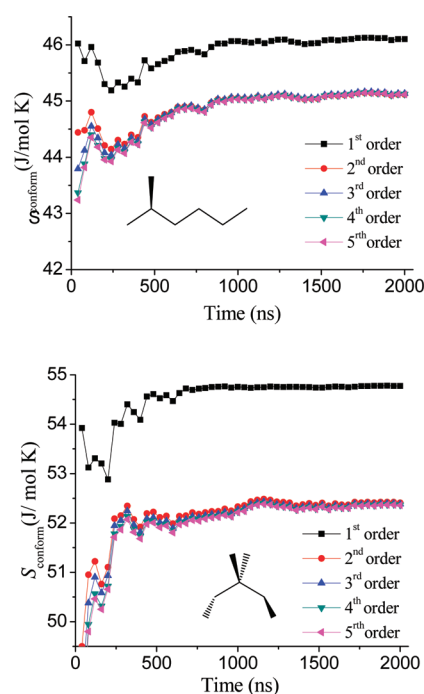


Figure 3. Convergence plots of the gas-phase conformational entropy (J/mol K) for 2-methyl-hexane and 3,3-dimethyl-pentane.

faster convergence was observed for the rest of the alkane molecules (see Figure S1 in the Supporting Information), and therefore, we conclude that the reported $S_{conform}$ values for these small alkane molecules essentially lack any statistical uncertainty. Although the

RRHO entropy is clearly the largest part, the S_{conform} values, which range between ~ 30 and ~ 50 J/(K mol), represent a significant contribution. Analyzing now the relative importance of correlation effects in S_{conform} , we find that the first-order MIE, which assumes that all torsion angles are independent variables, overestimates the conformational entropy. This behavior is not entirely unexpected because, for these small systems, the convergence plots in Figure 3 and Figure S1 (Supporting Information) strongly suggest that all correlation effects are effectively taken into account by our calculations, and as a consequence, the total entropy diminishes with respect to the first order value. However, we see in Figure 3 and Figure S1 that the entropy curves at first order recover most of the total conformational entropy; that is, entropy reduction due to correlation effects is rather small, only $1-5$ J/(K mol). Moreover, it turns out that this reduction is basically due to pair correlation, as the converged second order values are practically indistinguishable from the rest of the higher order S_{conform} entropies in all cases. Therefore, from these test calculations on the selected alkane molecules, we conclude that (a) S_{conform} is a significant entropic contribution and (b) the degree of correlation among the torsional degrees of freedom is rather low so that the second order MIE approximation gives sufficiently accurate values for the examined alkane molecules.

Before comparing our theoretical absolute entropies for the various C_6H_{14} and C_7H_{16} isomers with experimental data, it is necessary to realize that our S_{conform} calculations, which are based on classical MD simulations, imply that individual atoms in a covalently bound molecule are distinguishable particles. Therefore, in order to compare our data with experimental third-law entropies, we have to remove from the S_{conform} values the conformational entropy that arises from the number of possible rearrangements (Ω_{term}) that a single molecule can formally undergo through internal rotations about bonds to terminal symmetrical groups (e.g., $-\text{CH}_3$) without altering any molecular property. For instance, in the case of 2,2-dimethyl-butane, internal rotation of each of the four terminal methyl groups as well as around the C2–C3 bond generates three conformers per rotatable bond of identical energy and molecular properties, so that 2,2-dimethyl-butane has a total of $\Omega_{\text{term}} = 3^5$ possible intramolecular arrangements that would result in an entropy contribution $S_{\text{conform}}^{\text{term}} = R \ln \Omega_{\text{term}} = 45.67$ J/(K mol) (the $S_{\text{conform}}^{\text{term}}$ values for the rest of the alkane molecules can be likewise computed). The addition of the entropy differences, $\Delta S_{\text{conform}} = S_{\text{conform}} - S_{\text{conform}}^{\text{term}}$, to the RRHO mean values \bar{S} allow us to properly compare between theoretical and experimental data. Note also that entropic effects due to the presence of enantiomeric conformers are taken into account automatically by the S_{conform} calculations. We see in Figure 2 that the linearity of the correlation plot between the experimental and RRHO-based entropies is quite improved after having added the $\Delta S_{\text{conform}}$ term to \bar{S} . Thus, the squared correlation coefficient is now 0.96, and the intercept is around -2.6 J/(K mol).

The acceptable goodness of the linear fit between experimental and theoretical data in Figure 2 confirms the importance of S_{conform} and supports the usefulness of decomposing the entropy into its vibrational and conformational parts. In the Supporting Information, we present more calculations of the absolute entropies of the eight alkane molecules by using the above-mentioned “mixture of conformers” model.^{44,45} These calculations, which are based entirely on QM data, suggest that a significant fraction of the observed error in the RRHO&MIE conformational entropy calculations can emerge from small unbalances in the probability density functions of torsion angles. On the other hand, further

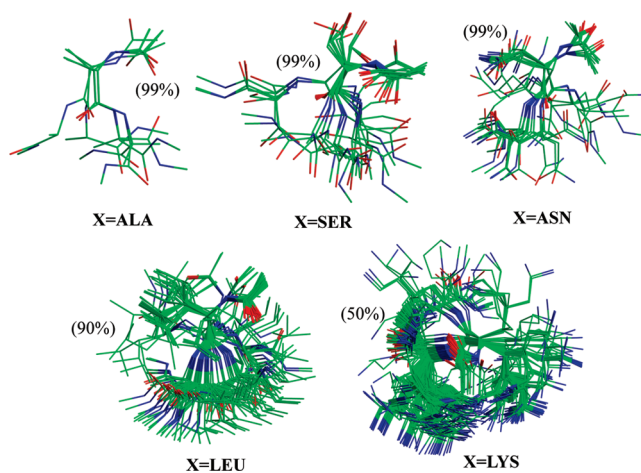


Figure 4. Superposition of the energetically distinguishable conformers of the Ace–X–X–Nme systems. The percentage of the MD snapshots represented by the conformers is given in parentheses.

support for applying the entropy decomposition can be found in previous works,^{44–49} following the “mixture of conformers” strategy. Therefore, it may also be relevant to comment here about the quality of the results of the “mixture of conformers” model. For example, the mean unsigned difference (MUD) between experimental and theoretical gas-phase entropies reported by Guthrie⁴⁷ for 128 organic compounds with up to 10 carbon atoms is 3.7 J/(K mol) (data derived from unscaled B3LYP/6-31G** frequencies and a semiempirical estimation of the ΔS_{mix} term). A similar MUD (4.4 J/(K mol)) was observed in our calculations on the eight alkane compounds.

Conformational Entropy of Dipeptides. The purpose of introducing the MIE approach within the context of the conformational entropy calculations is to capture correlation among torsional motions. For the alkane molecules studied in this work, it turns out that such correlation effects are almost negligible, and in any case, they are entirely accounted for by the pairwise approximation (i.e., second order correlation). Therefore, it is necessary to analyze more complex molecules in order to assess the ability of our approach for estimating higher order entropic contributions and its dependence on the dimensionality of the problem as defined by the number M of rotatable bonds. To this end, we examined five dipeptide molecules (Ace–X–X–Nme) of increasing size with X = Ala, Ser, Asn, Leu, and Lys (i.e., the peptides are capped by acetyl (Ace) and N-methyl (Nme) groups). The potential ability of these molecules to dynamically form and break intramolecular interactions through direct H-bond contacts or through-space electrostatic forces can introduce a significant correlation in their torsional motions and simultaneously maintain an important flexibility.

Figure 4 shows the superposition of the most populated conformers of the Ace–X–X–Nme systems that were obtained from minimizing 2000 equally spaced MD snapshots in each system and selecting the energetically distinguishable structures. As expected, all of the dipeptides are flexible molecules in the gas phase that experience frequent conformational transitions along the MD simulations. For the X = Ala system with $M = 8$ rotatable bonds, 99% of conformational variability is represented by only four structures, but the number of populated conformers grows up rapidly with M : the X = Lys system with $M = 16$ rotatable bonds populates more than 1200 different structures. Of course,

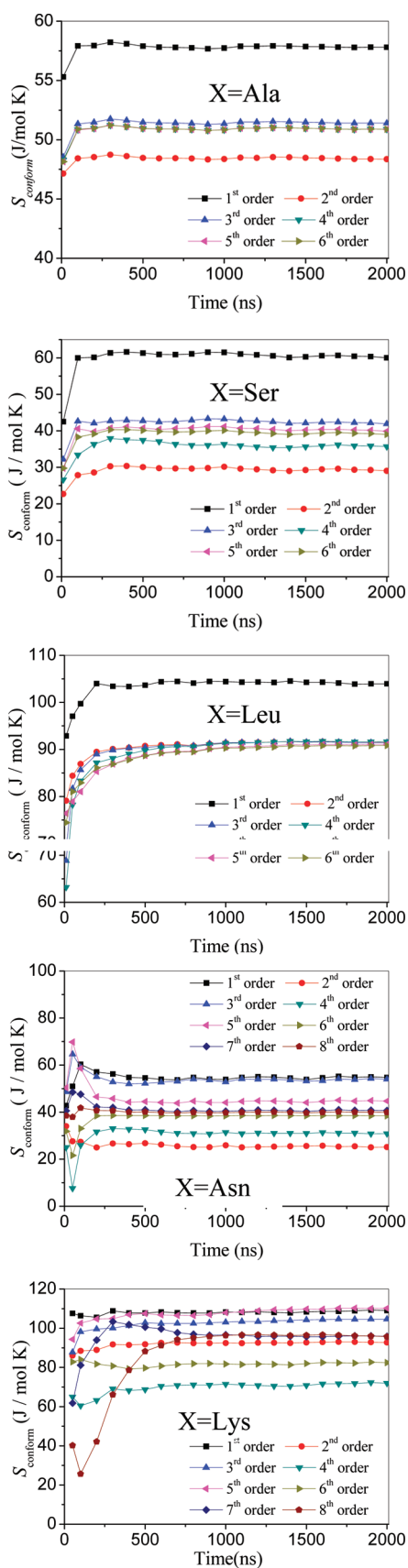


Figure 5. Convergence plots of the gas-phase conformational entropy ($J/(K \text{ mol})$) for the Ace–X–X–Nme molecules with X = Ala, Ser, Asn, Leu, and Lys.

the structure and relative abundance of the conformers are influenced by H-bond interactions that interconnect main chain groups (e.g., between the Ace–C=O carbonyl and the H–N–X amide at the second residue position) and/or side chain groups. Thus, we found that, in general, at least two or three H-bond contacts have abundances of about 55–65% so that they are quite stable interactions.

The conformational entropies at various orders of the Ace–X–X–Nme molecules are plotted along the simulation time in Figure 5. For the majority of the systems, it turns out that their S_{conform} profiles converge to nearly zero-slope curves after $\sim 0.5 \mu\text{s}$, thus suggesting that the MD simulations have exhaustively explored their phase space. However, the limiting S_{conform} values show a clear dependence on the expansion order n of the MIE method (see also Table 2). In consonance with expectations, the magnitude and relative weight of the n -order corrections increase with M on going from X = Ala ($M = 8$) to X = Lys ($M = 16$). Thus, with respect to the unimportance of high order correlation effects in Figure 3, we see now that both the intramolecular interactions and the partial rigidity introduced by the amidic bonds in the dipeptide molecules can have an impact on the conformational entropy thanks to the appearance of correlation among torsional motions. For the Ace–Ala–Ala–Nme system, which has only one more rotatable bond than the C_7H_{16} isomers and populates just a few conformers (see Figure 4), the second and third order corrections to S_{conform} are noticeable: -9.4 and $3.0 J/(K \text{ mol})$. These correlation effects are more clearly seen in the rest of the Ace–X–X–Nme systems as their S_{conform} curves span a range of tenths of a $J/(K \text{ mol})$. In order to achieve convergence in the S_{conform} values with respect to the expansion order n , MIE calculations up to sixth to eighth order were required for several systems (see Table 2). Moreover, we see in Figure 5 that n -order corrections to the conformational entropy fluctuate in both sign and magnitude depending on the molecular system, revealing thus the true complexity of the mutual information expansion. Nevertheless, the extended sampling (2×10^6 configurations) and the ability of the MIE calculations to estimate high order correlation effects allow us to obtain reasonably well-converged S_{conform} values for all of the Ace–X–X–Nme systems. In the case of the X = Asn and X = Lys systems, the difference between the S_{conform} estimations up to seventh and eighth order amounts to only $\sim 0.2 \text{ kJ/mol}$ in terms of free energies at 300 K. Similarly, the corresponding free energy differences for the S_{conform} values of the rest of the Ace–X–X–Nme dipeptides at the fifth and sixth orders are also negligible.

An interesting comparison can be made between the X = Leu ($M = 14$) and Asn ($M = 13$) systems. On one hand, the hydrophobic system has a considerable conformational entropy ($\sim 90 J/(K \text{ mol})$), which is in consonance with its relatively large number of accessible conformers. Curiously, high order ($n > 3$) corrections to S_{conform} are very small, which is very probably related to the fact that the Ace–Leu–Leu–Nme molecule forms only two intramolecular H-bond interactions with moderate abundances ($< 35\%$). On the other hand, much fewer conformers are populated during the MD trajectory of the X = Asn system, but correlation effects are now rather important, most likely because three to four H-bond interactions with abundances between 30% and 65% involving either side chain or backbone groups are constantly being formed and broken during the simulation, thus coupling the conformational changes of torsions separated by several covalent bonds. For the X = Lys system with $M = 16$ rotatable bonds, in addition to the relevance of correlation effects,

Table 2. Limiting Values of the Conformational Entropy at Various Orders (in J/(mol K)) for Dipeptides in the Gas Phase after 2.0 μ s of Simulation Time^a

	N_E	M	S_{conform}							
			1	2	3	4	5	6	7	8
Ace-(Ala) ₂ -Nme	14	8	57.80	48.36	51.40	50.88	50.88	50.93		
Ace-(Ser) ₂ -Nme	25	10	60.03	29.02	41.92	35.66	39.90	39.06		
Ace-(Asn) ₂ -Nme	69	13	54.76	25.14	54.00	30.86	44.70	38.41	40.66	39.90
Ace-(Leu) ₂ -Nme	232	14	103.96	91.49	91.44	91.64	91.10	90.84		
Ace-(Lys) ₂ -Nme	1357	16	109.08	92.70	104.64	71.97	110.18	82.41	95.91	95.74

^a The number of rotatable bonds considered in the conformational entropy calculations (M) and the number of energetically distinguishable conformers (N_E) observed in a sample of 2000 MD snapshots are also indicated. A total of 2×10^6 MD snapshots were used in all of the calculations.

its S_{conform} curves at high orders exhibit a slower convergence with respect to the simulation time (see Figure 5) as a consequence of the larger dimensionality of the problem, which demands a greater sampling effort to obtain reliable probability mass functions for all of the torsion angles and their combinations. Nevertheless, as mentioned above, the S_{conform} values for $X = \text{Lys}$ were converged to $\sim 96 \text{ J}/(\text{K mol})$ at orders $n = 7$ and 8, a value which happens accidentally to be quite close to the second order one (~ 93 ; see Table 2).

Absolute Entropy of the PFG Peptide. On the basis of the test calculations on the alkane molecules, it seems that there is no clear preference between the “mixture of conformers” and the conformational entropy frameworks for estimating the total entropy of small molecules. However, a sharp difference between the two formally equivalent approaches appears in larger biomolecules for which it is almost impossible to explore all their conformational space as required by the “mixture of conformers” approach. But as mentioned in the Introduction, the partitioning of the single molecule entropy, the discretization of the torsional degrees of freedom, and the adoption of the MIE technique can provide altogether a workaround to the sampling limitation for estimating the total entropy of complex molecular systems from the output provided by extensive MD simulations. To better illustrate the potential applicability of such an approach, we present here the results obtained for the PFG hexapeptide molecule, which is known to be a ligand of the MMP enzymes.

The 2.0 μ s MD trajectory of the PFG peptide in explicit solvent, which was started from the initial structure favored by the LMOD algorithm, populates two different conformational regions of the solute molecule characterized by radii of gyration of ~ 5.0 and $\sim 6.0 \text{ \AA}$, respectively (Figure S3 in the Supporting Information). Secondary structure analyses assign a helical conformation to the central (i.e., 2–6) residues in $\sim 55\%$ of the analyzed snapshots (we note in passing that the fact that PFG tends to adopt a helical structure seems in consonance with the ability of the MMPs to bind and hydrolyze collagen peptide chains that have also a helical structure). However, although PFG possesses some secondary structure, it still exhibits a relatively large dynamical flexibility through either its backbone or side chain motions, and therefore, it is conceivable that conformational entropy could be large enough to play a significant role in the free energy change upon binding of PFG to MMPs.³⁸

Most likely, the full theoretical understanding of the role played by entropy in the activity of PFG would require the computation of its absolute entropy in aqueous solution, as suggested by our previous calculations on the complexes formed between the MMP-2 enzyme and small peptide substrates.³⁸ In our approach,

such calculations could be achieved by combining the RRHO entropy of PFG using an implicit solvent model with the solute conformational entropy derived from the MD simulation so that the resulting entropy would be combined with other free energy terms as defined by approximate methodologies like the MM-PBSA protocol.⁵⁴ As the PFG hexapeptide contains 111 atoms and virtually all of the MD snapshots correspond to energetically distinguishable conformers, we decided to perform energy minimization and normal mode calculations on a subset of 10 000 MD snapshots using the NAB package and the AMBER03 force field (note that DFT calculations would be computationally too expensive). As shown in Figure 6A, although the RRHO entropies significantly fluctuate, the resulting time series over 2.0 μ s is rather stable and the corresponding mean value of S ($1415 \text{ J}/(\text{K mol})$) was estimated to within a standard error of only $0.3 \text{ J}/(\text{K mol})$. This small uncertainty suggests that the average RRHO entropy of PFG can be considered sufficiently converged for most purposes.

Turning our attention to the convergence plots of conformational entropy in Figure 6B derived from 10^6 MD snapshots, we note first that the S_{conform} values show a large dependence on the MIE order. In fact, the sum of marginal entropies leads to a limiting value of $137 \text{ J}/(\text{K mol})$ that is $\sim 60\%$ above the entropy estimations made from second to fifth order approximations. Given that PFG is a flexible molecule that contains a significant number of rotatable bonds ($M = 25$) and has several polar groups capable of forming H-bond interactions, the entropy reduction caused by correlation effects is well understood. However, for the same reasons, the S_{conform} calculations at the fourth and fifth orders now have a poor convergence with respect to simulation time: there remains an uncertainty of a few $\text{J}/(\text{K mol})$ in their limiting values after 2.0 μ s. Besides the sampling limitations at high order, it is also clear that the S_{conform} values of PFG have a non-negligible uncertainty with regard to the n order employed in the calculations. For example, the limiting value of S_{conform} at second order is $\sim 3 \text{ kJ}/\text{mol}$ in terms of free energies below that at third order.

Segregation of the conformational entropy into backbone and side chain contributions allows us to further analyze the origin of the large correlation effects in the dynamics of PFG. We see in Figures 6C,D that the backbone ($M = 11$) and side chain ($M = 14$) S_{conform} curves show an acceptable convergence at the various orders. Curiously, the entropy curves reflecting the conformational changes of the ψ and ϕ torsion angles present marked oscillations and converge more slowly than those of the amino acid side chains, indicating thus that the solvent-exposed side chains move faster than the backbone chain, and therefore, their motions are more efficiently sampled by the MD simulations. Correlation

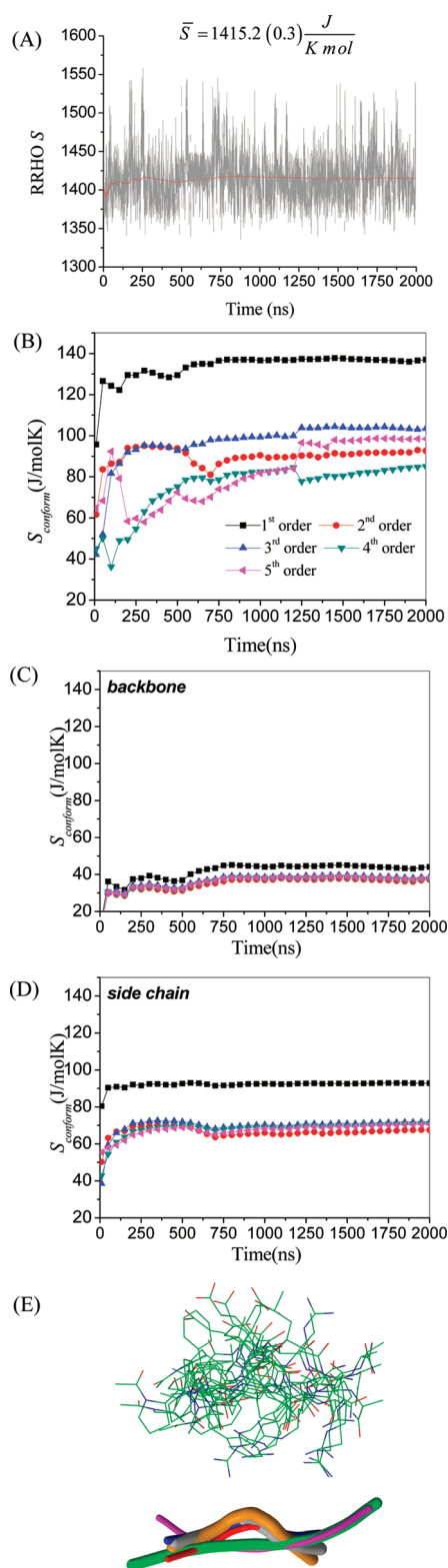


Figure 6. (A) RRHO entropy (in J/(K mol)) calculated for 10 000 snapshots extracted at 200 ps intervals from the 2.0 μ s MD simulations of PFG. The average value and its standard error in parentheses are also indicated. (B–D) Convergence plots of the PFG conformational entropy (in J/(K mol)). (E) Superposition of the most populated representative structures derived from clustering analyses both in wire-frame and ribbon model representation. Thickness of the ribbon models corresponds to the number of snapshots represented by each model.

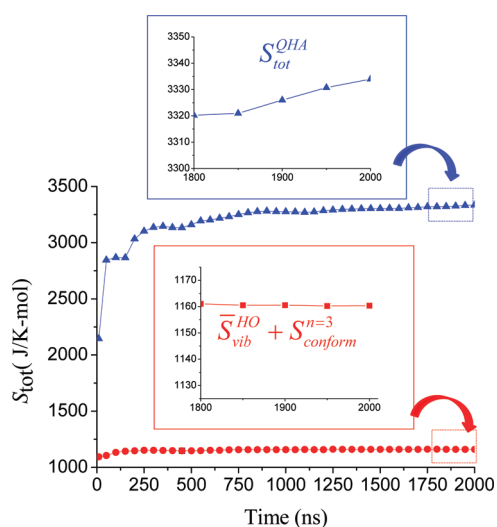


Figure 7. Convergence plots of the total entropy of PFG (excluding translation and rotation) estimated by the QHA method and the addition of the mean vibrational entropy and the conformational entropy at third order.

among the ψ and ϕ angles is only moderate (~ 6 J/(K mol)) and is largely captured by the second order approximation. Similarly, second- or third-order corrections to $S_{conform}$ are able to include most of the correlation effects due to the conformational motions of the side chains, but the magnitude of the concomitant entropy reduction is larger (~ 22 J/(K mol)). When comparing the plots in Figure 6B–D, it is clear that the source of the stronger correlation effects in the conformational entropy of the PFG peptide stem from the coupling between the backbone and side chain motions. We also see that, in terms of their convergence properties, the separate $S_{conform}$ values for the backbone and side-chain torsions are much more reliable than the global data as the segregated entropy plots reach stable plateaus and the free energy differences between the third-, fourth-, and fifth-order estimations are very small (~ 0.1 kJ/mol in terms of free energies).

Finally, we compare the additive entropy $\bar{S}_{vib}^{HO} + S_{conform}^{n=3}$ with the results provided by the quasi-harmonic approximation QHA, which constructs a pseudo-Hessian matrix directly from the covariance matrix of the Cartesian coordinates. To remove the overall translation of the center of mass and the overall rotation of the protein, the 2×10^6 MD snapshots employed in the QHA calculations were superposed on top of each other using a least-squares fit. As mentioned in the Introduction, the QHA method exhibits several disadvantages (e.g., neglecting supralinear correlations, approximating multimodal distributions to a unimodal one, etc.) that ultimately result in a large nonsystematic overestimation of S_{config} . This effect is clearly observed in Figure 7 as the QHA calculations lead to a very large entropy value (~ 3333 J/(K mol)). Although the entropy decomposition should give an upper limit to the true entropy, the $\bar{S}_{vib}^{HO} + S_{conform}^{n=3}$ value (~ 1160 J/(K mol)) is much lower than the QHA one even though the $-TS_{conform}$ contribution has a statistical uncertainty of several kJ/mol. Moreover, besides the larger overestimation of S_{config} the QHA calculations also exhibit worse convergence properties (see the two insets in Figure 7).

From the results of the test calculations on PFG, it can be concluded that the total conformational entropy has a considerable weight (10%) in its single-molecule entropy and that the effects of dynamic correlations among torsional angles are far

from being small. These calculations point out that both under-sampling and poor convergence with respect to the MIE order are two closely related problems of the S_{conform} calculations that need to be assessed in practical applications. However, although the uncertainty in the total S_{conform} of PFG can have an impact of several kJ/mol on free energy, we believe that meaningful results can be achieved from partially converged entropy curves like those shown in Figures 6B–D for PFG. For example, computation of entropy differences (e.g., upon peptide binding to a host molecule) could benefit from partial cancellation of errors and provide approximately constant ΔS values at different MIE orders.^{22,38} Alternatively, consideration of a subset of torsion angles (e.g., backbone ψ and ϕ angles) could be enough for capturing the relative change in entropy occurring in peptide folding or molecular association processes. In any case, an intensive MD sampling followed by the estimation of conformational entropy differences including correlation effects would be required to fully understand the thermodynamical forces controlling many biomolecular processes that alter the conformational dynamics of the involved molecules. In this respect, the present test calculations support the ability of our approach for including high order correlation effects that have been neglected so far in most of the previous studies using nonparametric methods for estimating the configurational entropy.

SUMMARY AND CONCLUSIONS

In this work, we have pursued the implementation of the partitioning of the intramolecular entropy into vibrational and conformational contributions as originally proposed by Karplus et al., in order to estimate the absolute entropy of single biomolecules from MD simulations. In our approach, a key element consists of the characterization of the conformational state of a given molecule by means of the discretization of the time evolution of its torsional motions about single bonds. This process leads naturally to the computation of the conformational entropy as Shannon entropy, which is subsequently added to the average translational–rotational–vibrational entropy computed by means of normal model calculations on a series of MD snapshots. To help overcome sampling limitations in the computation of the conformational probability mass function of large molecules, we use the mutual information expansion, which has also been employed in other entropy methods, to systematically include correlation effects among torsion angles. Although this protocol has been developed for treatment with relatively large systems, it must be emphasized that its core assumption, that is, the combination of the RRHO entropy with the conformational entropy, is formally equivalent to the “mixture of conformers” strategy that has been routinely used to predict absolute entropies of small molecules with good accuracy.

On the basis of the different test calculations that have been presented in this work, we can draw the following conclusions regarding the applicability and/or reliability of our approach: (a) The gross of the absolute entropy is computed with the RRHO approximation, which constitutes a straightforward computational protocol that avoids the need of discriminating between stiff or soft degrees of freedom. (b) From a quantitative point of view, the combination of the RRHO entropy with the conformational (or mixing) entropy yields results that are quite close to experimental data, as shown by our calculations on the alkane molecules or by other results previously reported in the literature. (c) For computing reliable conformational entropies of small- or

medium-sized molecules that exhibit a rich dynamical behavior, it is essential to capture correlation effects arising from the coupling of torsional motions through intramolecular interactions, and in this respect, the use of the MIE method together with the discretization of the torsional motions constitutes an interesting alternative capable of computing high order corrections quite efficiently. (d) Segregation of conformational entropy into different components (e.g., backbone and side chain terms) can be easily implemented, thus revealing the origin and relative importance of correlation effects.

Finally, it is also important to comment on some limitations of the RRHO-conformational entropy calculations on the basis of the results obtained for the PFG polypeptide. Thus, it is clear that the RRHO&MIE calculations for this kind of system may demand a considerable amount of computer time, particularly if energy minimizations and normal mode calculations have to be carried out on systems containing thousands of atoms, even though an MM method was employed. Similarly, the computation of conformational entropy corrections at high orders ($n > 5-8$) can be rather expensive too. In addition, application of the present approach to large molecules can suffer from convergence issues with respect to the simulation time needed to extract the probable mass functions of individual torsions and groups of torsions, and with reference to the MIE order that is required to capture correlation effects. We believe that the calculation of relative conformational entropies rather than absolute ones using low order approximations and/or for a subset of torsions may benefit from error cancellation. However, the problems of the RRHO&MIE technique should be mitigated by increasing the MD sampling, which in turn, is becoming more and more accessible thanks to the continuous improvement in the efficiency in computer hardware and simulation algorithms. Moreover, the fact that the current protocol has been shown to exhibit a much better convergence behavior than the QHA method in the case of the polypeptide entropy calculation is a promising result that should stimulate further methodological and computational experimentation aimed at overcoming the computational bottlenecks and/or convergence problems.

ASSOCIATED CONTENT

S Supporting Information. Derivation of the entropy partitioning (eq 8). Mathematical proof of the equivalence between eqs 9 and 10. Figure S1 showing convergence plots of conformational entropies for various alkane molecules. Figure S2 and Tables S1 and S2 summarizing the results of the “mixtures of conformers” entropy calculations on the alkane molecules. Figure S3 showing the time evolution of the PFG radius of gyration. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +34-985103689. Fax: +34-985103125. E-mail: dimas@uniovi.es.

ACKNOWLEDGMENT

This research was supported by the following grants: FICyT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. E.S. thanks MEC for his FPU contract.

REFERENCES

- (1) Carlsson, J.; Aqvist, J. Calculations of Solute and Solvent Entropies from Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385.
- (2) Brady, G. P.; Sharp, K. E. Entropy in Protein Folding and in Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215.
- (3) Fitter, J. A. Measure of Conformational Entropy Change During Thermal Protein Unfolding Using Neutron Spectroscopy. *Biophys. J.* **2003**, *84*, 3924.
- (4) Bachmann, A.; Kiefhaber, T.; Boudko, S.; Engel, J.; Bächinger, H. P. Collagen Triple-Helix Formation in All-Trans Chains Proceeds by a Nucleation Growth Mechanism with a Purely Entropic Barrier. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13897.
- (5) Creamer, T. P.; Rose, G. D. Side-Chain Entropy Opposes α -Helix Formation but Rationalizes Experimentally Determined Helix-Forming Propensities. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5937.
- (6) Chapagain, P. P.; Parra, J. L.; Gerstman, B. S.; Liu, Y. Sampling of States for Estimating the Folding Funnel Entropy and Energy Landscape of a Model α -Helical Hairpin Peptide. *J. Chem. Phys.* **2007**, *127*, 075103.
- (7) Choa, S. S.; Levya, Y.; Wolynesa, P. G. Quantitative Criteria for Native Energetic Heterogeneity Influences in the Prediction of Protein Folding Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 434.
- (8) Schäfer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on a Reversibly Folding Peptide: Changes in Solute Free Energy Cannot Explain Folding Behavior. *Proteins: Struct., Funct., Genet.* **2001**, *56*, 43.
- (9) Grünberg, R.; Nilges, M.; Leckner, J. Flexibility and Conformational Entropy in Protein-Protein Binding. *Structure* **2006**, *14*, 683.
- (10) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational Entropy Changes Upon Lactose Binding to the Carbohydrate Recognition Domain of Galectin-3. *J. Biomol. NMR* **2009**, *45*, 157.
- (11) Stone, M. J. Nmr Relaxation Studies of the Role of Conformational Entropy in Protein Stability and Ligand Binding. *Acc. Chem. Res.* **2001**, *34*, 379.
- (12) Chang, C. A.; Chen, C.; Gilson, M. K. Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534.
- (13) Baron, R.; McCammon, J. A. (Thermo)Dynamic Role of Receptor Flexibility, Entropy, and Motional Correlation in Protein-Ligand Binding. *ChemPhysChem* **2008**, *9*, 983.
- (14) Killian, B. J.; Yudenfreund-Kravitz, J.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an Hiv-Derived Ptp Nonapeptide. *J. Mol. Biol.* **2009**, *389*, 315.
- (15) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325.
- (16) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* **1984**, *51*, 1011.
- (17) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617.
- (18) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289.
- (19) Baron, R.; Gunsteren, W. F. v.; Hünenberger, P. H. Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation. *Trends Phys. Chem.* **2006**, *11*, 87.
- (20) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* **2006**, *28*, 655.
- (21) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations Via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127*.
- (22) Suárez, E.; Díaz, N.; Suárez, D. Entropic Control of the Relative Stability of Triple-Helical Collagen Peptide Models. *J. Phys. Chem. B* **2008**, *112*, 15248.
- (23) Li, D.-W.; Bruschiweiler, R. In Silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102*, 118108.
- (24) Hensen, U.; Grubmüller, H.; Lange, O. F. Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional Configuration Spaces. *Phys. Rev. E* **2009**, *80*, 011913.
- (25) Hensen, U.; Lange, O. F.; Grubmüller, H. Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS ONE* **2010**, *5*, e9179.
- (26) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer-Verlag: Berlin, 2007.
- (27) Meirovitch, H. Methods for Calculating the Absolute Entropy and Free Energy of Biological Sys. *J. Mol. Recognit.* **2010**, *2*, 153.
- (28) Meirovitch, H.; Chelvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and Their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Protein Pept. Sci* **2009**, *10*, 229.
- (29) Schäfer, H.; Mark, A. E.; Gunsteren, W. F. v. Absolute Entropies from Molecular Dynamics Simulation Trajectories. *J. Chem. Phys.* **2000**, *113*, 7809.
- (30) Chang, C.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017.
- (31) Cover, T. M.; Thomas, J. C. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2006.
- (32) Di Nola, A.; Berendsen, H. J. C.; Edholm, O. Free Energy Determination of Polypeptide Conformations Generated by Molecular Dynamics. *Macromolecules* **1984**, *17*, 2044.
- (33) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5*, 3150.
- (34) Numata, J.; Wan, M.; Knapp, E. Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation. *Genome Inform.* **2007**, *18*, 192.
- (35) Gorla, M. N.; Leonenko, N. N.; Mergel, V. V.; Novi-Inverardi, P. L. A New Class of Random Vector Entropy Estimators and Its Applications in Testing Statistical Hypotheses. *J. Nonparametr. Stat.* **2005**, *17*, 277.
- (36) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29*, 1605.
- (37) Ohkubo, Y. K.; Thorpe, I. F. Evaluating the Conformational Entropy of Macromolecules Using an Energy Decomposition Approach. *J. Chem. Phys.* **2006**, *124*, 024910.
- (38) Díaz, N.; Suarez, D.; Suarez, E. Kinetic and Binding Effects in Peptide Substrate Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and Qm/Mm Calculations. *Proteins* **2010**, *78*, 1.
- (39) Karplus, M.; Ichiye, T.; Pettit, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083.
- (40) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092.
- (41) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000.
- (42) Matsuda, H. Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration. *Phys. Rev. E* **2000**, *62*, 3098.
- (43) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, 1964.
- (44) DeTar, D. F. Theoretical Ab Initio Calculation of Entropy, Heat Capacity, and Heat Content. *J. Phys. Chem. A* **1998**, *102*, 5128.
- (45) DeTar, D. F. Calculation of Entropy and Heat Capacity of Organic Compounds in the Gas Phase. Evaluation of a Consistent Method without Adjustable Parameters. Applications to Hydrocarbons. *J. Phys. Chem. A* **2007**, *111*, 4464.
- (46) Block, D. A.; Armstrong, D. A.; Rauk, A. Gas Phase Free Energies of Formation and Free Energies of Solution of Rc-Centered Free Radicals from Alcohols: A Quantum Mechanical-Monte Carlo Study. *J. Phys. Chem. A* **1999**, *103*, 3562.

- (47) Guthrie, J. P. Use of Dft Methods for the Calculation of the Entropy of Gas Phase Organic Molecules: An Examination of the Quality of Results from a Simple Approach. *J. Phys. Chem. A* **2001**, *105*, 8495.
- (48) Bouchoux, G.; Bimbong, R. G.-B.; Nacer, F. Gas-Phase Protonation Thermochemistry of Glutamic Acid. *J. Phys. Chem. A* **2009**, *113*, 6666.
- (49) Bouchoux, G.; Bourcier, S.; Blanc, V.; Desaphy, S. Gas Phase Protonation Thermochemistry of Phenylalanine and Tyrosine. *J. Phys. Chem. B* **2009**, *113*, 5549.
- (50) East, A. L. L.; Radom, L. Ab Initio Statistical Thermodynamical Models for the Computation of Third-Law Entropies. *J. Chem. Phys.* **1997**, *106*, 6665.
- (51) Brown, R. A.; Case, D. A. Second Derivatives in Generalized Born Theory. *J. Comput. Chem.* **2006**, *27*, 1662.
- (52) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree–Fock, Møller–Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100*, 16502.
- (53) Johnson, R. D.; Irikura, K. K.; Kacker, R. N.; Kessel, R. Scaling Factors and Uncertainties for Ab Initio Anharmonic Vibrational Frequencies. *J. Chem. Theory Comput.* **2010**, *6*, 2822.
- (54) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: Mm-Pb(Gb)Sa Studies on the Protein–Protein Complex Ras–Raf. *J. Comput. Chem.* **2003**, *25*, 238.
- (55) Lee, M. S.; Olson, M. A. Calculation of Absolute Protein–Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* **2009**, *90*, 864.
- (56) Taylor, C. C. Automatic Bandwidth Selection for Circular Density Estimation. *Comput. Stat. Data An.* **2008**, *52*, 3493.
- (57) Suárez, E.; Suárez, D. Manuscript in preparation.
- (58) Case, D. A.; Darden, T. A.; Cheatham, I. T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, K. F.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER*; 10th ed.; University of California: San Francisco, 2008.
- (59) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668.
- (60) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *14*, 1999.
- (61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, C.02 ed.; Gaussian Inc.: Wallingford, CT, 2004.
- (62) Turk, B. E.; Huang, L. L.; Cantley, L. C. Determination of Protease Cleavage Site Motifs Using Mixture-Based Oriented Peptide Libraries. *Nat. Biotechnol.* **2001**, *19*, 661.
- (63) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011.
- (64) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401.
- (65) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- (66) Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.
- (67) *Modeling Unusual Nucleic Acid Structures*; Macke, T., Case, D. A., Eds.; American Chemical Society: Washington, DC, 1998.
- (68) Scott, D. W. Correlation of the Chemical Thermodynamic Properties of Alkane Hydrocarbons. *J. Chem. Phys.* **1974**, *60*, 3144.
- (69) Kilpatrick, J. E.; Pitzer, K. S. The Thermodynamics of 2,2-Dimethylbutane, Including the Heat Capacity, Heats of Transition, Fusion and Vaporization and the Entropy. *J. Am. Chem. Soc.* **1946**, *68*, 1066.
- (70) Finke, H. L.; Messerly, J. F. 3-Methylpentane and 3-Methylheptane: Low-Temperature Thermodynamic Properties. *J. Chem. Thermodyn.* **1973**, *5*, 247.
- (71) Huffman, H. M.; Gross, M. E.; Scott, D. W.; McCullough, J. P. Low Temperature Thermodynamic Properties of Six Isomeric Heptanes. *J. Phys. Chem.* **1961**, *65*, 495.

SUPPORTING INFORMATION

Entropy Calculations of Single Molecules by
Combining the Rigid–Rotor and Harmonic–Oscillator
Approximations with Conformational Entropy
Estimations from Molecular Dynamics Simulations

*Ernesto Suárez, Natalia Díaz, and Dimas Suárez**

dimas@uniovi.es

A) ENTROPY DECOMPOSITION

Herein we assume that the translational and rotational degrees of freedom of a molecular system have been removed (*e.g.*, by fitting all the configurations to a reference structure) and the resulting potential energy surface in terms of the remaining internal degrees of freedom can be approximated by a set of *distinguishable* energy basins as shown in the Scheme 1 of the Manuscript. Each j -basin, which can have an arbitrary shape, is unambiguously assigned to a molecular conformer. We also assume that the quantization of the potential energy of each conformer results in a series of discrete *vibrational* energy levels ($i=0, 1, \dots$). Therefore, the configurational state of a molecule is given by an energy level, E_{ij} , which can be expressed with respect to an arbitrary reference. The total entropy associated with the intramolecular degrees of freedom can be computed using the Shannon expression as follows:

$$S_{tot} = -R \sum_{i,j} p\{v=i, c=j\} \ln p\{v=i, c=j\} \quad (\text{S.1})$$

where $p\{v=i, c=j\}$ is the probability that the molecular system occupies the i -vibrational level of the j -conformer. In the canonical ensemble, this probability can be expressed in terms of the canonical partition functions of the whole ensemble:

$$p\{v=i, c=j\} = \frac{\exp(-\beta E_{ij})}{\sum_{i,j} \exp(-\beta E_{ij})} = \frac{\exp(-\beta E_{ij})}{\sum_j Q_j} \quad (\text{S.2})$$

In this expression, we identify the sum over the vibrational states with the vibrational partition function Q_j of each conformer. We assume at this point that the vibrational entropy is a *local* property of each energy basin, that is, it is associated to one and only one energy well. Thus, the vibrational entropy of the j conformational state can be expressed as:

$$S_{vib}^j = -R \sum_i p\{v=i|c=j\} \ln p\{v=i|c=j\} \quad (\text{S.3})$$

where $p\{v = i|c = j\}$ is the conditional probability of the i -vibrational state conditioned by $c = j$.

Again this probability can be expressed in terms of the Boltzmann equation as:

$$p\{v = i|c = j\} = \frac{\exp(-\beta E_{ij})}{\sum_i \exp(-\beta E_{ij})} = \frac{\exp(-\beta E_{ij})}{Q_j} \quad (\text{S.4})$$

where Q_j is now the partition function calculated over the j -conformational state. On the basis of

the Bayes theorem, the statistical weight of each conformer $p\{c = j\}$ should be equal to:

$$p\{c = j\} = \frac{p\{v = i, c = j\}}{p\{v = i|c = j\}} = \frac{Q_j}{\sum_j Q_j} \quad (\text{S.5})$$

The statistical thermodynamic meaning of $p\{c = j\}$ is more clearly expressed in terms of the partition functions Q_j , which represent the non-normalized statistical weight of the j -conformers.

By combining Eqs (S.1) and (S.5), we obtain:

$$S_{tot} = -R \sum_{i,j} [p_j p\{v = i|c = j\}] \ln [p_j p\{v = i|c = j\}] \quad (\text{S.6})$$

where p_j stands for the probability of the conformational state j (*i.e.*, $p\{c = j\}$). By expanding

the logarithm of the product, reordering terms, and realizing that $\sum_i p\{v = i|c = j\} = 1$, we have:

$$S_{tot} = -R \sum_j p_j \left(\sum_i p\{v = i|c = j\} \ln p\{v = i|c = j\} \right) - R \sum_j p_j \ln p_j$$

where we can easily identify the vibrational and conformational entropy terms with the inner sum

in the first term and the sum in the second term, respectively. Hence, we can write up the

following entropy decomposition formula, which is equivalent to Eq. (S.1):

$$S_{tot} = \sum_j p_j S_{vib}^j + S_{conform}$$

B) REFORMULATION OF THE MIE EXPRESSION

PROPOSITION 1. *The usual truncation of the mutual information expansion for the estimation of the total entropy of a system composed of M ensembles (i.e., subsystems) including correlation effects up to n -order with $n \leq M$ is ^{S1}*

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}), \quad (\text{S.7})$$

where $\mathcal{J} = \{\mathcal{J}_1, \dots, \mathcal{J}_k\}$ runs over all possible subsets of $\mathcal{A} = \{A_1, \dots, A_M\}$ with k elements and the mutual information I_k shared among k ensembles is expressed in terms of the subsystem entropies as

$$I_k(\mathcal{J}) = \sum_{l=1}^k (-1)^{l+1} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=l}} S(\mathcal{I}). \quad (\text{S.8})$$

We can affirm that equation (S.7) is equal to

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.9})$$

Proof. To proof the equivalence between (S.7) and (S.9), we must verify that the equality (S.7)=(S.9) holds for $n=1$, because this particular case will be the *seed* for a formal proof of the general case by means of mathematical induction. According to the induction principle, a statement $P(n)$ is true $\forall n \in \mathbb{N}$ if the following two statements hold true:

1. The statement is true for the first element ($n=0$ or $n=1$), in our case: $n=1$.
2. If the statement is true for an arbitrary n then it is true for $(n+1)$

Let us first show that (S.7)=(S.9) is valid for $n=1$. Notice that, from the definition of I_k in (S.8), it follows that for $k=1$ the set $\mathcal{J} = \{\mathcal{J}_1\}$ has only one element, which is ultimately one of the i -elements of \mathcal{A} , i.e., $\mathcal{J} = \{A_i\}$. Since $\{A_i\}$ itself, is the only subset of $\{A_i\}$ with cardinality one, then $I_1(A_i) = S(A_i)$. Thus, we have that:

$$S^{(1)}(\mathcal{A}) = \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=1}} S(\mathcal{J}) = \sum_{A_i \in \{A_1, \dots, A_M\}} S(A_i) = \sum_i S(A_i). \quad (\text{S.10})$$

Next we obtain the same result from (S.9). Note that if $n=1$ in (S.9), then $i=0$ and the (S.9) equation is transformed to

$$S^{(1)}(\mathcal{A}) = \binom{M-1}{0} \sum_{\substack{\mathcal{I} \subset \{A_1, \dots, A_M\} \\ |\mathcal{I}|=1}} S(\mathcal{I}),$$

where by definition $\binom{M-1}{0} = \frac{(M-1)!}{((M-1)-0)!0!} = 1$. Finally, we simplify the notation of the single sum in the latter expression:

$$S^{(1)}(\mathcal{A}) = \sum_{\substack{\mathcal{I} \subset \{A_1, \dots, A_M\} \\ |\mathcal{I}|=1}} S(\mathcal{I}) = \sum_{A_i \in \{A_1, \dots, A_M\}} S(A_i) = \sum_i S(A_i),$$

obtaining thus the same result as in (S.10). Therefore, the statement (S.7)=(S.9) is true for $n=1$.

The second step in our formal proof of (S.7)=(S.9) is to verify that, if the statement is true for an arbitrary n , then it is true for $(n+1)$. Thus, particularizing the equation (S.9) for $n+1$:

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^{n+1} \left[\sum_{i=0}^{n+1-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.11})$$

Next we split the outer sum by extracting the last term ($k = n+1$):

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n+1-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}) \quad (\text{S.12})$$

Similarly, we split the middle sum in the first term for $i = n-k+1$:

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} + (-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}) \quad (\text{S.13})$$

Under the hypothesis that equation (S.9) is true for n , the last equation can be rewritten as

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + \sum_{k=1}^n \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}), \quad (\text{S.14})$$

where the two separate sums can be regrouped, obtaining

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + \sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.15})$$

Following analogous steps, expression (S.7) can be particularized for $n+1$ as

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^{n+1} (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}), \quad (\text{S.16})$$

and then transformed into

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + (-1)^n \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} I_{n+1}(\mathcal{J}). \quad (\text{S.17})$$

By comparing (S.15) and (S.17), it is clear that the proof of the original proposition implies that the following equality must hold true:

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = (-1)^n \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} I_{n+1}(\mathcal{J}). \quad (\text{S.18})$$

Therefore we must prove equality (S.18). First, we express I_{n+1} in the right side of (S.18) in terms of the subsystem entropies (for convenience, the l -index used in definition (S.8) is replaced now by k):

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} \sum_{k=1}^{n+1} (-1)^{n+k+1} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k}} S(\mathcal{I})$$

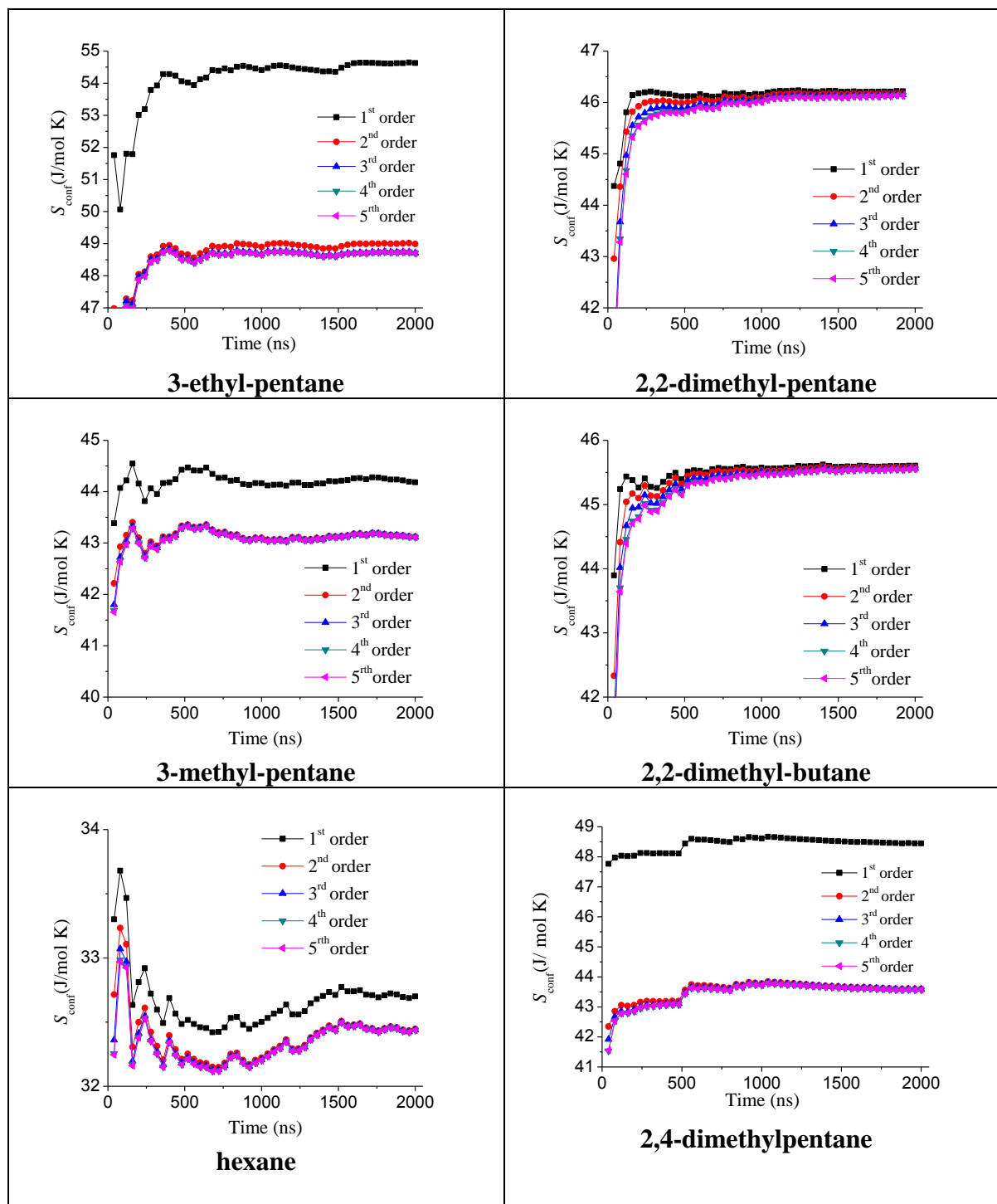
Reordering the sums in the right side and knowing that $(-1)^{n-k+1} = (-1)^{n+k+1}$:

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = \sum_{k=1}^{n+1} (-1)^{n-k+1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k}} S(\mathcal{I}).$$

Finally, to prove that this last expression is true, we transform the double sum over the subsets \mathcal{J} and \mathcal{I} in the right side into a single sum. To this end we need to count the number of times a given subset \mathcal{I} appears while summing over \mathcal{J} . Once the k elements of \mathcal{I} are selected, there are $(n+1)-k$ unselected elements of \mathcal{J} . Thus, the number of times a given \mathcal{I} will appear is the number of possibilities for selecting $(n+1)-k$ elements from the rest $M-k$ elements, which is exactly $\binom{M-k}{n-k+1}$. Therefore, expression (11) is true, thereby demonstrating the validity of the

Proposition 1.

Figure S1. Convergence plots of the gas-phase conformational entropy ($J/mol K$) for different alkane molecules.



C) ENTROPY CALCULATIONS ON ALKANE MOLECULES USING THE “MIXTURES OF CONFORMERS” APPROACH

Although addition of the QM RRHO entropies to the $S_{conform}$ values obtained from classical MD simulations gives reasonable absolute entropies, we decided to investigate the source of the remaining error that can have values up to $\sim 9 J(K mol)$ in the case of hexane. On one hand, we believe that the level of theory for the harmonic frequency calculations, B3LYP/cc-pVTZ, is unlikely to account for these residual errors because the entropy of 2,2-dimethyl-butane, which exists as a single conformer, is matched quite closely by the B3LYP RRHO calculations (the error is below $1 J(K mol)$). However, the GAFF parameters, which have been designed for reproducing molecular properties in condensed phase, could introduce some bias in the probability density functions of the torsional degrees of freedom in the gas-phase, affecting thus to the quality of the results. To further explore this possibility, we estimated the absolute entropies of the eight alkane molecules by using the “mixture of conformers” model. Thus, we computed first the B3LYP/cc-pVTZ free energies of the energetically-distinguishable conformers for each alkane compound. Subsequently, the number of *distinct* conformers for each compound was determined by manual inspection in order to take into account the existence of enantiomeric conformers^{S2} and other non-quiral conformers of equal energy (note that this tedious task is performed automatically by the MD-based conformational calculations). Then the relative population of all the conformers was obtained by means of the Maxwell-Boltzmann distribution in terms of the relative free energies (without $S_{conform}$). The combination of the Maxwell-Boltzmann-averaged RRHO and the entropy of mixing term based on the number of distinct conformers, which are collected in Table S1, leads to quite similar data to those reported in Table 1, the MUD between experimental and computational entropies being now $2.6 J(K mol)$ as

compared with the 4.4 $J(K mol)$ of the MIE calculations based on the MD probability density functions.

In principle, the calculations performed from the “mixture of conformers” viewpoint at the B3LYP/cc-pVTZ level suggest that the performance of the general purpose MM force fields is not far from that of purely DFT methodologies in predicting conformational entropies, but they do not entirely clarify the source of the remaining errors in the entropy calculations. Thus, we examined the influence of high-level correlated methods in the relative population of the alkane conformers. To this end we computed the MP2 energies of the hexane conformers extrapolated to their complete basis set limit (CBS)^{S3} by combining single-point MP2/cc-pVQZ and MP2/cc-pV5Z correlation energies according to the Schwartz extrapolation scheme (HF/cc-pV5Z energies were used here as CBS HF energies). Similarly, we also computed the single-point CCSD(T)/cc-pVTZ energies. Then the corresponding CCSD(T)/CBS values were estimated by means of the following “composite” formula:

$$E_{CCSD(T) CBS} \approx E_{CCSD(T)/cc-pVTZ} + (E_{MP2/CBS} - E_{MP2/cc-pVTZ})$$

The MP2 calculations were performed with the TURBOMOLE program package^{S4} in the framework of the “resolution-of-the-identity” approximation (RI-MP2) using the appropriate auxiliary basis set^{S5} while the CCSD(T) energies were obtained with MOLPRO.^{S6} We replaced the B3LYP/cc-pVTZ electronic energies with the composite ab initio values in the Gibbs energies, and the Maxwell-Boltzmann population of each hexane conformer was then recomputed. Interestingly, the S_{mix} of hexane based on the ab initio data is now 20.73 $J(K mol)$, 3.99 $J(K mol)$ above that derived from the DFT data (see Table S1). This result is due to the fact that high level correlated methods tend to stabilize the more compact conformers of hexane with

respect to the fully extended conformer, whose relative population decreases from 55.1% (DFT) to 37.4% (ab initio). The predicted value for the absolute entropy of hexane, 386.98 $J/(K mol)$ has now an error of only 1.8 ($J/ K mol$). Overall, we conclude that a significant fraction of the observed error in our conformational entropy calculations can emerge from small unbalances in the probability density functions of torsion angles. Nonetheless, other issues like the treatment of anharmonicity effects, hindered torsions, non-ideal gas behavior correction factors, experimental uncertainties, etc, should probably be considered for further accuracy.^{S7-8}

Figure S2 Correlation plot between experimental and theoretical (“mixture of components”) absolute entropies (in $J/(K mol)$) for the 8 alkane molecules considered in this work.

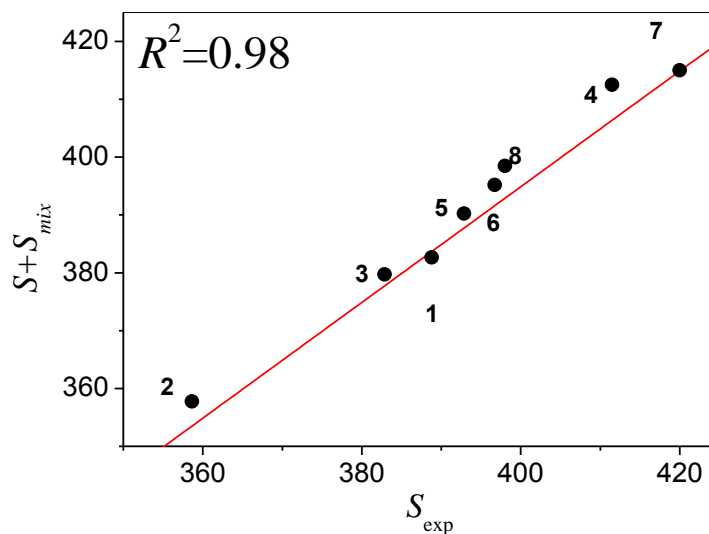


Table S1. Entropy contributions (in $J/mol\ K$) at 298 K according to the “mixture of conformers” protocol in which the conformational population distribution is estimated by the Maxwell–Boltzmann formula in terms of B3LYP/cc-pVTZ free energies.

	Alkane	S_{exp}	$\bar{S}^{(1)}$	S_{mix}	$\bar{S} + S_{mix}$
1	hexane ⁽²⁾	388.82	365.93 (366.25)	16.74 (20.73)	382.67 (386.98)
2	2,2-dimethyl-butane	358.65	357.77	0.00	357.77
3	3-methyl-pentane	382.88	365.75	13.97	379.72
4	3-ethyl-pentane	411.50	391.68	20.82	412.50
5	2,2-dimethyl-pentane	392.88	388.53	1.73	390.26
6	2,4-dimethyl-pentane	396.73	386.56	8.64	395.20
7	2-methyl-hexane	419.99	397.49	17.51	415.0
8	3,3-dimethyl-pentane	398.02	382.21	16.27	398.48

$$^{(1)}\bar{S} = S_{trans} + \bar{S}_{rot}^{RR} + \bar{S}_{vib}^{HO}$$

and using the B3LYP/cc-pVTZ geometries and frequencies. The S_{rot} and S_{vib}

contributions are averaged according the relative abundance of the distinct conformers.

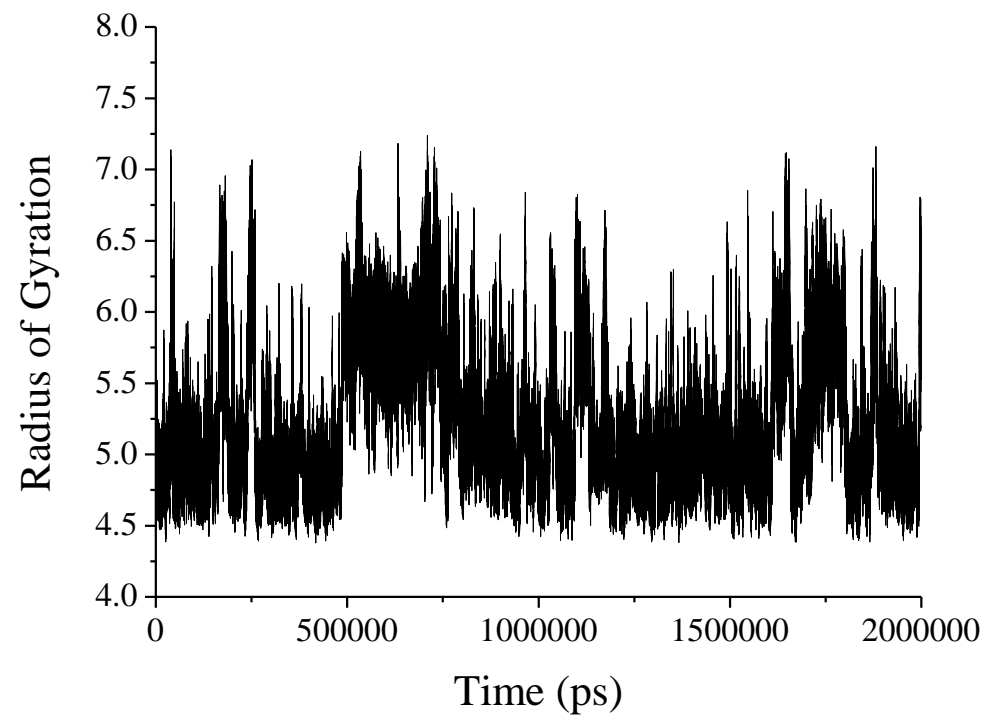
⁽²⁾Data in parentheses correspond to the values obtained from high level ab initio calculations (see text for details).

Table S2. Relative conformational energies (*kJ/mol*) of the 9 hexane conformers considered at this work at different levels of theory. Geometries were optimized at the B3LYP/cc-pVTZ level. Thermal contributions to Gibbs free energy were obtained from B3LYP/cc-pVTZ frequency calculations.

	B3LYP/cc-pVTZ	G^{therm}	G B3LYP/cc-pVTZ	MP2/VTZ	MP2/CBS	CCSD(T)/ cc-pVTZ	G Composite ⁽¹⁾
1 (all <i>trans</i>)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	3.78	1.04	4.82	2.42	2.51	2.52	3.65
3	7.38	1.06	8.44	4.05	4.15	4.38	5.54
4	3.72	1.23	4.95	2.43	2.49	2.52	3.81
5	7.52	1.21	8.73	4.84	4.97	5.01	6.35
6	7.94	0.71	8.65	5.37	5.47	5.46	6.28
7	14.16	1.34	15.50	11.45	11.63	11.31	12.83
8	14.41	1.06	15.47	11.96	12.24	11.73	13.08
9	11.05	1.42	12.47	5.64	5.76	6.18	7.72

$$(1) G_{\text{composite}} = E_{\text{CCSD(T)/VTZ}} + (E_{\text{MP2/CBS}} - E_{\text{MP2/VTZ}}) + G_{\text{B3LYP/VTZ}}^{\text{therm}}$$

Figure S3. Radius of gyration (\AA) along the MD simulation for the **PFG** peptide.



REFERENCES

- (S1) Matsuda, H. Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration. *Phys. Rev. E* **2000**, *62*, 3098.
- (S2) Nasipuri, D. *Stereochemistry of Organic Compounds : Principles and Applications* 4th ed.; New Academic Science Kent, UK, 2011.
- (S3) Martin, J. M. L. In *Nato Asi Symposium Volume Asic 535, Energetics of Stable Molecules and Reactive Intermediates.*; Minas da Piedade, M. E., Ed.; Kluwer Academic Publishers: Dordrecht, 1999, p 373.
- (S4) Ahlrichs, R.; Bär, M.; Baron, H. P.; Rüdiger, B.; Böcker, S.; Crawford, N.; Deglmann, P.; Ehrig, M.; Eichkorn, K.; Elliot, S.; Furche, F.; Haase, F.; Häser, M.; Horn, H.; Hättig, C.; Huber, C.; Huniar, U.; Kattannek, M.; Köhn, A.; Kölmel, C.; Kollwitz, M.; May, K.; Nava, P.; Ochsenfeld, C.; Öhm, H.; Patzelt, H.; Rappoport, D.; Rubner, O.; Schäfer, A.; Schneider, U.; Sierka, M.; Treutler, O.; Unterreiner, B.; von Arnim, M.; Weigend, F.; Weis, P.; Weiss, H. In *Turbomole*; 5.9 ed. Karlsruhe (Germany), 2005.
- (S5) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. Ri-Mp2: Optimized Auxiliary Basis Sets and Demonstration of Efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (S6) Werner, H. J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, K.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. In *MOLPRO* Cardiff, UK, 2008.
- (S7) Lin, C. Y.; Izgorodina, E. I.; Coote, M. L. How Accurate Are Approximate Methods for Evaluating Partition Functions for Hindered Internal Rotations? *J. Phys. Chem. A* **2008**, *112*, 1956.
- (S8) Ellingson, B. A.; Lynch, V. A.; Mielke, S. L.; Truhlar, D. G. Statistical Thermodynamics of Bond Torsional Modes: Tests of Separable, Almost-Separable, and Improved Pitzer–Gwinn Approximations. *J. Chem. Phys.* **2006**, *125*, 084305.

***2.1.1.4 Distinguishability in Entropy Calculations: Chemical Reactions,
Conformational and Residual Entropies***

Ernesto Suárez

Entropy. **2011**, 13, 1533-1540

Commentary

Distinguishability in Entropy Calculations: Chemical Reactions, Conformational and Residual Entropy

Ernesto Suárez

Departamento de Química Física y Analítica, Universidad de Oviedo, Julián Clavería 8, 33006, Oviedo, Spain; E-Mail: ernesto@fluor.quimica.uniovi.es; Tel.: +34-985103492; Fax: +34-985103125

Received: 15 June 2011; in revised form: 2 August 2011 / Accepted: 20 August 2011 /

Published: 23 August 2011

Abstract: By analyzing different examples of practical entropy calculations and using concepts such as conformational and residual entropies, I show herein that experimental calorimetric entropies of single molecules can be theoretically reproduced considering chemically identical atoms either as distinguishable or indistinguishable particles. The broadly used correction in entropy calculations due to the symmetry number and particle indistinguishability is not mandatory, as an *ad hoc* correction, to obtain accurate values of absolute and relative entropies. It is shown that, for *any* chemical reaction of *any* kind, considering distinguishability or indistinguishability among identical atoms is irrelevant as long as we act consistently in the calculation of all the required entropy contributions.

Keywords: particle distinguishability; chemical reactions; conformational entropy; residual entropy; Gibbs paradox

1. Introduction

In the statistical treatment of a system of N identical particles, it is customary to divide the partition function by $N!$ in order to avoid the overcounting of states due to particle indistinguishability [1]. Analogously, the single molecule partition function is divided by the external symmetry number σ_{ext} that corresponds to the number of indistinguishable molecular orientations, and by the internal symmetry number σ_{int} that accounts for the number of indistinguishable conformers (see [2] for a convincing discussion on symmetry numbers). In all cases, the reduction in microstates is related to the concept of indistinguishability: since chemically identical atoms are considered as indistinguishable, any permutation among them would lead to the same state.

The solution of the so-called *Gibbs paradox* [3–8], is probably the most famous example where the same kind of correction has been applied. Gibbs proposed an *ad hoc* reduction in the entropy of an N -particle system by the amount $-k_B \ln N!$, where k_B is Boltzmann's constant. This entropy diminution, which corresponds to a reduction in the number of microstates accessible to the system by the permutation symmetry number $N!$, was able to correct what Gibbs considered as an unphysical situation, that is, the fact that entropy increases after mixing two (identical) ideal gases both being initially at the same temperature and pressure.

The unphysical situation that Gibbs tried to avoid in mixing processes is consistent with the concept of entropy as extensive property as held by Gibbs himself. However, it may be interesting to remember that the thermodynamic definition of entropy proposed by Clausius in 1865 does not reveal anything about how the entropy behaves as the number of particles N changes. The Clausius definition only allows us to compute the difference in entropy between two thermodynamics states of a *closed system*. Pauli noticed this incompleteness and showed what additional condition must be imposed in order to define an extensive entropy, suggesting that entropy, as defined by Clausius, is not intrinsically an extensive property [4,9]. In any case, the extensivity and the indistinguishability are concepts closely connected to each other in the context of the Gibbs arguments, which have been supported and rejected more than once in an ongoing debate [3,4,7,10,11].

From the point of view of classical statistics, whenever identical particles are distinguishable (Maxwell–Boltzmann statistics), it turns out that the entropy reduction by a term of $-k_B N \ln \sigma$ is in contrast with the idea of a magnitude that grows with the number of microscopic complexions compatible with the macroscopic state of the system [6], because ultimately, the result of any symmetry operation including any permutation is another microscopic complexion. Nevertheless, it is well known that the experimental 3^{rd} law entropies of small molecules can be reproduced with extreme accuracy if the entropy reduction is employed. It therefore *appears* that the experimental values can be reproduced only by using this correction, or equivalently, by adopting truly quantum statistics from which the entropy reduction emerges naturally due to the symmetry of the wave function [12].

The residual entropy [13–15] is another concept that can be linked to the indistinguishability. When we assert that a perfect crystal at $0K$ has null entropy, we are implicitly assuming from the statistical standpoint that any permutation of two identical particles does not lead to a new microstate. Although the residual entropy is only relevant when is empirically detectable [13,14] and can be related to a potentially measurable latent heat [15], the concept will be helpful when we analyse absolute entropies in the context of distinguishable particles. Because in this scenario, the residual entropy would be present even if a reversible path to the solid state at $0K$ were available.

Herein, through the careful analysis of various practical cases, I support the idea that considering identical atoms as indistinguishable particles is not mandatory in order to compute entropy values that are in agreement with experiment [10,16,17]. I show that classical treatment (distinguishable particles) can reproduce experimental entropy values without the need for any adjustment due to *weaknesses* of the classical model. All that is required is to be consistent with all the implications arising from distinguishability, including the consequences for the residual and conformational entropy (if any) of the involved molecules. I also show with two examples the innocuous effect of the distinguishability

on the entropy change in chemical reactions obtaining for all cases the same result as that obtained by considering identical atoms as indistinguishable.

The entropy of mixing and the Gibbs paradox, however, is out of the scope of this work because the problem has recently been solved for distinguishable particles without any *ad hoc* correction [8,17]. In this respect, the present work tries to generalize the idea to any other chemical transformation of any kind, where symmetry changes might take place. The implications of these ideas could be particularly relevant for approximate calculations of absolute entropies in which quantum mechanical and classical statistics are mixed in order to estimate different entropic contributions.

2. Discussion

2.1. Indistinguishable Particles and Third Law Entropies

Nowadays, 3rd law entropies of small molecules in the gas-phase can be computed easily and with remarkable precision by feeding thermodynamic statistical formulae with molecular properties computed with quantum chemical methods [18]. For example, Table 1 shows both theoretical and experimental entropy values reported in the literature for small alkanes [19]. For simplicity, the examples in Table 1 are selected so that the ω possible conformers for each molecule, if any, are not only isoenergetic, but also chemically identical. Thus, the conformational entropy would be zero, depending whether or not we are considering indistinguishability or distinguishability among identical conformers.

Table 1. B3LYP/cc-pVTZ theoretical and experimental entropies in $JK^{-1}mol^{-1}$ [19].

Molecule	σ_{ext}	Theory	Experiment	Abs. Error
methane	12	186.20	186.37	0.17
ethane	6	228.50	229.16	0.66
propane	2	270.20	270.31	0.11
methylpropane	3	295.50	295.70	0.20
dimethylpropane	12	306.74	306.00	0.74
2,2-dimethylbutane	1	358.70	358.40	0.30

Since molar entropies are being dealing with, the number of particles is chosen to be the Avogadro Number (N_a), expressing the entropy corrections preferably in terms of the gas constant $R = k_B N_a$. The theoretical values in Table 1 are Rigid-Rotor Harmonic-Oscillator (RRHO) entropies obtained from standard statistical thermodynamic formulae at the B3LYP/cc-pVTZ level of theory, where B3LYP is a hybrid density functional and cc-pVTZ denotes the correlation consistent basis set used [20]. Standard formulae refers to the fact that in all cases the reported theoretical entropies are reduced due to the symmetry including the permutation symmetry (*i.e.*, reduced by the terms $-R \ln \sigma_{ext}$ and $-k_B \ln N_a!$) [12]. In principle, we should also correct the entropy due to the internal symmetry number $\sigma_{int} = \omega$ by adding $-R \ln \omega$. However, it is well known that RRHO entropies do not capture all the intramolecular entropy, as they lack the purely conformational part of the entropy [21–23], which is in our case exactly $R \ln \omega$ and, therefore, the last correction is automatically done due to the deficiencies of the RRHO

method. As can be seen in Table 1, the theoretical results are, without any doubt, in good agreement with the experimental values.

2.2. Distinguishable Particles and Third Law Entropies

If the particles are distinguishable, the entropy correction is not justified and there are new entropy terms that should be taken into account. The conformational entropy, for instance, is now not canceled and consequently the corresponding term $R \ln \omega$, as well as the one due to the external symmetry $R \ln \sigma_{ext}$, must be added to each of the theoretical values in Table 1. By doing so, the agreement of the theoretical data with the experimental values *apparently* worsens. However, we realise that standard experimental calorimetric entropies are ultimately an entropy change from $T = 0K$ to $T = 298K$. This change is equal to the absolute entropy if the 3rd law holds, *i.e.*,

$$S_{T=298K} = S_{T=0K} + \int_{T=0K}^{T=298K} \frac{\delta Q}{T} \quad (1)$$

To interpret the experimental results assuming distinguishable particles, it can be noted first that, in the examples, any *formal* conformational change of a single molecule near $0K$, as well as any rotational symmetry operation, will lead to a different microscopic complexion compatible with the macroscopic state [13]. Therefore, a residual entropy should be considered for these molecules having a value of $R \ln (\omega \sigma_{ext})$. This quantity must be added to the original experimental values and the resulting entropy values, which assume particle distinguishability, maintain the agreement between theory and experiment (see Table 2).

Table 2. Theoretical (B3LYP/cc-pVTZ) and experimental entropy values in $JK^{-1}mol^{-1}$ augmented by $R \ln (\omega \sigma_{ext})$ due to the distinguishability.

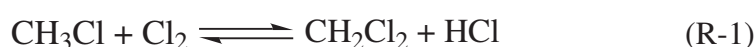
Molecule	ω	σ_{ext}	Theory	Experiment	Abs. Error
methane	1	12	206.86	207.03	0.17
ethane	3	6	252.53	253.19	0.66
propane	3 ²	2	294.23	294.34	0.11
methylpropane	3 ³	3	332.03	332.23	0.20
dimethylpropane	3 ⁴	12	363.93	363.19	0.74
2,2-dimethylbutane	3 ⁵	1	404.37	404.07	0.30

At this point the reader might wonder why, if all particles are taken as distinguishable, the uncertainty due to the permutation symmetry in the solid state at $0K$ has not been considered. After all, any permutation would give a new *different* microstate. In fact, it could have been done, but it would have changed nothing, because in such a case the term $k_B \ln N_a!$ needs to also be added to the theoretical value because the translational part of the entropy is computed in its corrected form $S_t = R(\ln q_t + \frac{5}{2})$, where q_t is the translational partition function [1]. Note that the corrected form is conceptually equivalent to the “reduced” entropy used by Cheng [17]. In general any other intra- or extra-molecular permutation between identical but distinguishable atoms can both be considered in the gas phase and in the solid state nearby $0K$, and the agreement between theory and experiment would be unaffected.

2.3. Entropy Changes in Chemical Reactions: Is There Any Difference?

The statement that the absolute entropy of a system depends on a subjective decision, to consider or not that identical atoms or particles are indistinguishable, most likely seems awkward. It is no less subjective, however, than setting an arbitrary reference in order to transform a relative magnitude into an absolute one. There are an infinite number of possible functions that would give the correct experimental entropy change, and therefore they all meet the original Clausius thermodynamic definition. The entropy change is the magnitude that must be invariant regardless of any considerations. Through two simple examples, both points of view discussed above (considering identical particles distinguishable or not) will be shown as totally equivalent.

Let us first consider the following equilibrium reactions in the gas phase:



If a quantum chemical program is used to optimize the molecular geometries, carry out the corresponding frequency calculations and compute the RRHO entropies without considering any symmetry operation except the identity, the entropy values, say, $S_{\text{CH}_3\text{Cl}}^{\text{RRHO}}$, $S_{\text{Cl}_2}^{\text{RRHO}}$, $S_{\text{CH}_2\text{Cl}_2}^{\text{RRHO}}$, and $S_{\text{HCl}}^{\text{RRHO}}$ would be obtained. For convenience the required entropy corrections are introduced explicitly, then the estimated entropy change in (R-1) is

$$\begin{aligned} \Delta S &= \Delta S_{\text{nosym}}^{\text{RRHO}} + \Delta(-R \ln \sigma_{\text{ext}}) \\ &= \Delta S_{\text{nosym}}^{\text{RRHO}} - R \ln \frac{\sigma_{\text{ext}}(\text{CH}_2\text{Cl}_2)\sigma_{\text{ext}}(\text{HCl})}{\sigma_{\text{ext}}(\text{CH}_3\text{Cl})\sigma_{\text{ext}}(\text{Cl}_2)} \\ &= \Delta S_{\text{nosym}}^{\text{RRHO}} + R \ln 3 \end{aligned} \quad (2)$$

where $\Delta S_{\text{nosym}}^{\text{RRHO}} = S_{\text{CH}_2\text{Cl}_2}^{\text{RRHO}} + S_{\text{HCl}}^{\text{RRHO}} - S_{\text{CH}_3\text{Cl}}^{\text{RRHO}} - S_{\text{Cl}_2}^{\text{RRHO}}$. In principle, ΔS would reproduce the experimental entropy change provided that the level of theory in the calculations is adequate.

Considering distinguishability in the same reaction (R-1), it is now obvious that the CH_3Cl molecule can be formed by any of the 3 distinguishable Cl atoms involved, furthermore, the three numerable H atoms can be reordered in two different forms not superimposable by rotations, being the atoms in Cl_2 completely determined by our first selection. Hence, the reactants would have an additional uncertainty that contributes to the entropy in $R \ln(2 \times 3)$. On the other hand, the CH_2Cl_2 molecule can be formed by any two of the three Cl and any two of the three H atoms, and once selected, there are two possible arrangements to be chosen between. Note that the atoms are numerable and we could obtain two enantiomeric configurations. The atoms in HCl will be determined once again by the previous selection and finally the entropy estimation under this new formalism is

$$\begin{aligned} \Delta S &= \Delta S_{\text{nosym}}^{\text{RRHO}} + R \ln \left\{ 2 \binom{3}{2} \binom{3}{2} \right\} - R \ln(2 \times 3) \\ &= \Delta S_{\text{nosym}}^{\text{RRHO}} + R \ln 18 - R \ln 6 \\ &= \Delta S_{\text{nosym}}^{\text{RRHO}} + R \ln 3 \end{aligned} \quad (3)$$

which is exactly the same result obtained above.

Let us consider a more complex example where the conformational entropy is also involved. In (R-2), the symmetry number is three for the methylpropane, one in the methylcyclopropane and two for the H, being $\Delta(-R \ln \sigma_{ext}) = R \ln (3/2)$.



Considering identical atoms as indistinguishable, there is no conformational entropy either in the methylpropane or in the methylcyclopropane molecules. The entropy change involved is $\Delta S = \Delta S_{nosym}^{RRHO} + R \ln (3/2)$.

If, on the contrary, identical atoms are distinguishable, the H atoms can be arranged in multiple different ways and the entropy value is not lower due to symmetry, but higher. Additionally, the conformational entropy must be taken into account since any conformational change in any methyl group would give a new different conformer. The uncertainty due to the arrangements of the carbon atoms (excluding the connectivity) is the same in reactant and products and will not be considered.

In order to build the reactant molecule (methylpropane), 10 H atoms need to be distributed into 4 “boxes” of capacities 3, 3, 3 and 1, where, in the boxes of capacity 3 (methyl groups), there are two possible enantiomeric arrangements. Also, each methyl group will contribute to the conformational entropy with 3 conformers, being the total number of complexions

$$\frac{10!}{3!3!3!1!} \times 2^3 \times 3^3 = 10!$$

For the products (methylcyclopropane and H₂) the carbon atoms which will close the cycle are selected first (there are $\binom{3}{2}$ possibilities), then we have 10 H atoms for 5 boxes of capacities 3, 2, 2, 1 and 2, where we included the H₂ molecule as the last box. Once again the methyl groups as well as the –CH₂– groups have two possible arrangements and each methyl group generates three different conformers. As a consequence, the total number of complexions is

$$\binom{3}{2} \times \frac{10!}{3!2!2!2!1!} \times 2^3 \times 3 = \frac{3}{2} \times 10!$$

and therefore the computed entropy is

$$\begin{aligned} \Delta S &= \Delta S_{nosym}^{RRHO} + R \ln (10! \times (3/2)) - R \ln 10! \\ &= \Delta S_{nosym}^{RRHO} + R \ln (3/2) \end{aligned} \quad (4)$$

obtaining again the same result under both formalisms.

However, two examples do not equate to a formal proof, the idea needs to be extended to *any* chemical reaction. To this end, notice that for distinguishable particles, those permutations that lead to a different arrangement, *i.e.*, not superimposable with the original one by rigid rotations are being considered. For a given system, a systematic way to compute the required number of permutations would be to consider all the possible permutations and then reduce this value taking into account the total symmetry number. For example, it is known that there are only two possible arrangements of the distinguishable atoms in the CH₄ molecule, this quantity is equal to the number of permutations of the H atoms (4!) divided by the symmetry number of a tetrahedral molecule ($\sigma = 12$).

In general, the number of permutations not superimposable by rotations (internal or external) of a system that have n_1 atoms of type 1, n_2 atoms of type 2, and so on, is equal to

$$\frac{\prod_i n_i!}{\prod_j \sigma_j} \quad (5)$$

where the denominator is the product of all the symmetry numbers of the system (reactants or products). If, for instance, the last expression in the reaction (R-1) is applied, it results in $(3!3!1!)/(2 \times 3) = 6$ and $(3!3!1!)/(2 \times 3) = 18$ complexions for the reactants and products respectively, the same results obtained above (see Equation (3)).

For a general reaction $React \rightleftharpoons Prod$, since the number and type of atoms is conserved, the numerator in (5) always cancels out in the difference $R \ln \frac{\prod_i n_i!}{\prod_j \sigma_{j,prod}} - R \ln \frac{\prod_i n_i!}{\prod_j \sigma_{j,react}}$, where $\sigma_{j,react}$ and $\sigma_{j,prod}$ are respectively the symmetry number of reactants and products. Consequently, the effect is equivalent to correcting the entropy change by $\Delta(-R \ln \sigma)$, being $\sigma = \prod_j \sigma_j$ the total symmetry number on each side of the chemical reaction. Note that the correction is the same as that for indistinguishable particles except for one point; in the above two examples when indistinguishable particles were considered, only external symmetry numbers were used, not because the internal symmetry was not present, but simply because of the flaws of the RRHO approach taken as “reference”. In other words, considering whether identical atoms are distinguishable or not has no effect on the entropy change in chemical reactions.

3. Conclusions

It has been explicitly shown through practical examples that, for all practical applications, it is irrelevant to consider indistinguishable particles or not in entropy calculations. The classical statistical treatment (distinguishable particles) is equally valid provided that the new degrees of freedom involved are taken into account properly. These arguments could be of particular interest for computing entropies in biochemical reactions where the classical treatment is ubiquitous. Even though in such systems it is quite common to observe chemical reactions like binding processes where no symmetry change takes place [21,22], care must be taken, because as we have seen, even under a classical formalism the symmetry should be considered for a proper entropy estimation.

Acknowledgements

The author thanks Dimas Suárez and Ramón López (Universidad de Oviedo) for their careful reading of the manuscript and their suggestions.

References

1. McQuarrie, D. *Statistical Mechanics*; University Science Books: Sausalito, CA, USA, 2000.
2. Gilson, M.K.; Irikura, K.K. Symmetry numbers for rigid, flexible, and fluxional molecules: Theory and applications. *J. Phys. Chem. B* **2010**, *114*, 16304–16317.
3. Ben-Naim, A. On the so-called Gibbs paradox, and on the real paradox. *Entropy* **2007**, *9*, 132–136.
4. Jaynes, E. The Gibbs paradox. In *Maximum Entropy and Bayesian Methods*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1992.

5. Lesk, A. On the Gibbs paradox: What does indistinguishability really mean? *J. Phys. A: Math. Gen.* **1980**, *13*, L111–L114.
6. Lin, S.K. Correlation of entropy with similarity and symmetry. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 367–376.
7. Lin, S.K. Gibbs paradox and the concepts of information, symmetry, similarity and their relationship. *Entropy* **2008**, *10*, 1–5.
8. Versteegh, M.A.M.; Dieks, D. The Gibbs paradox and the distinguishability of identical particles. *Am. J. Phys.* **2011**, *79*, 741–746.
9. Pauli, W. Thermodynamics and the Kinetic Theory of Gases. In *Pauli Lectures on Physics*; MIT Press: Cambridge, MA, USA, 1973.
10. Swendsen, R. Statistical mechanics of classical systems with distinguishable particles. *J. Stat. Phys.* **2002**, *107*, 1143–1166.
11. Nagle, J. Regarding the entropy of distinguishable particles. *J. Stat. Phys.* **2004**, *117*, 1047–1062.
12. Pathria, R. *Statistical Mechanics*; Butterworth-Heinemann: Oxford, UK, 1996.
13. Pauling, L. The structure and entropy of ice and of other crystals with some randomness of atomic arrangement. *J. Am. Chem. Soc.* **1935**, *57*, 2680–2684.
14. Kozliak, E. Consistent application of the Boltzmann distribution to residual entropy in crystals. *J. Chem. Educ.* **2007**, *84*, 493–498.
15. Kozliak, E.; Lambert, F.L. Residual entropy, the third law and latent heat. *Entropy* **2008**, *10*, 274–284.
16. Ercolany, G.; Piguet, C.; Borkovec, M.; Hamacek, J. Symmetry numbers and statistical factors in self-assembly and multivalency. *J. Phys. Chem. B* **2007**, *111*, 12195–12203.
17. Cheng, C.H. Thermodynamics of the system of distinguishable particles. *Entropy* **2009**, *11*, 326–333.
18. DeTar, D.F. Theoretical ab initio calculation of entropy, heat capacity, and heat content. *J. Phys. Chem. A* **1998**, *102*, 5128–5141.
19. NIST Computational Chemistry Comparison and Benchmark Database. Available online: <http://cccbdb.nist.gov/> (accessed on 10 June 2011).
20. Dunning, T. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.
21. Chang, C.; Chen, C.; Gilson, M.K. Ligand configurational entropy and protein binding. *Proc. Nat. Acad. Sci. USA* **2007**, *104*, 1534–1539.
22. Suárez, E.; Díaz, N.; Suárez, D. Entropic control of the relative stability of triple-helical collagen peptide models. *J. Phys. Chem. B* **2008**, *112*, 15248–15255.
23. Zhou, H.X.; Gilson, M.K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **2009**, *109*, 4092–4107.

2.1.1.5 *Multibody Local Approximation for Conformational Entropy*

Calculations on Biomolecules

Ernesto Suárez and Dimas Suárez

(Enviado para su publicación a *J. Chem. Theory Comput.*)

**Multibody Local Approximation: Application to
Conformational Entropy Calculations on Biomolecules**

Journal:	<i>Journal of Chemical Theory and Computation</i>
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Suárez, Ernesto; Universidad de Oviedo, Química Física y Analítica Suarez, Dimas; Universidad de Oviedo, Química Física y Analítica

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules

Ernesto Suárez and Dimas Suárez*

Departamento de Química Física y Analítica. Universidad de Oviedo. C/ Julián

Clavería, 8. 33006, Oviedo. Spain.

E-mail: ernesto@fluor.quimica.uniovi.es

TELEPHONE: +34-985103492

FAX: +34-985103125

ABSTRACT

Multibody type expansions like Mutual Information Expansions are widely used for computing or analyzing properties of large composite systems. The power of such expansions stems from their generality. Their weaknesses, however, are the large computational cost of including high order terms due to the combinatorial explosion and the fact that truncation errors do not decrease strictly with the expansion order. Herein, we take advantage of the redundancy of multibody expansions in order to derive an efficient reformulation that captures implicitly all-order correlation effects within a given cutoff, avoiding the combinatory explosion. This approach, which is cutoff dependent rather than order dependent, keeps the generality of the original expansions and simultaneously mitigates their limitations provided that a reasonable cutoff can be used. An application of particular interest can be the computation of the conformational entropy of flexible peptide molecules from Molecular Dynamics trajectories. By combining the multibody local estimations of conformational entropy with average values of the rigid-rotor & harmonic-oscillator entropic contributions, we obtain by far a tighter upper bound of the absolute entropy than the one obtained by the broadly used quasi-harmonic method.

INTRODUCTION

The well known Multibody Expansion (MBE) for energy estimation,¹⁻² the Mutual Information Expansion (MIE) for entropy calculations,³⁻⁴ the Molecular Tailoring Approach (MTA) for electronic densities,⁵ and other closely related methods,⁵⁻¹⁰ can all be considered as applications of a more general principle according to which every computable property or function S of any composite system $\mathcal{A} = \{A_1, \dots, A_M\}$, can be calculated without any approximation by the expansion

$$S(\mathcal{A}) = \sum_i S_i + \sum_{i < j} \{S_{ij} - S_i - S_j\} + \sum_{i < j < k} \{S_{ijk} - S_{ij} - S_{ik} - S_{jk} + S_i + S_j + S_k\} + \dots, \quad (1)$$

being $S_{ij\dots}$, the function S evaluated over a given subset $\mathcal{J} = \{A_i, A_j, \dots\}$ of \mathcal{A} . The same expression can be as

$$S(\mathcal{A}) = \sum_{k=1}^M \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} \sum_{m=1}^k (-1)^{k+m} \sum_{\substack{\mathcal{J} \subset \mathcal{I} \\ |\mathcal{J}|=m}} S(\mathcal{J}), \quad (2)$$

where the notation $|\cdot|$ is used for the cardinality (*i.e.* the number of elements of a set). The subset \mathcal{I} is formally an index running over all the subsets of \mathcal{A} with cardinality k , while \mathcal{J} runs over the subsets of \mathcal{I} with cardinality m . A formal proof of the generality of Eq.(2), which bears close resemblance to the inclusion-exclusion principle from set theory, is shown in the Proposition 2 of the Supporting Information (SI). In practical applications, however, $S(\mathcal{A})$ is approximated by a truncated form $S^{(n)}$ of Eq.(2) that takes into account the interactions up to a given order n . Due

to the combinatorial nature of this expansion, only low values of n (e.g., $n=1-3$) are computationally feasible.^{3,11-12}

For the assessment of configurational and/or conformational entropies, the truncation of Eq.(2) is usually written in terms of the so-called mutual information functions (MIFs) as follows

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} I_k(\mathcal{I}), \quad (3)$$

being I_k the mutual information shared among k subsystems, given by

$$I_k(A_{i_1}, \dots, A_{i_k}) = \sum_{m=1}^k (-1)^{m+1} \sum_{\substack{\mathcal{J} \subset \{A_{i_1}, \dots, A_{i_k}\} \\ |\mathcal{J}|=m}} S(\mathcal{J}), \quad (4)$$

where for a particular m , the sum over \mathcal{J} runs over all possible $\binom{k}{m}$ combinations^{4,13}

Regardless of the large computational cost of the expansion at high orders in Eq.(3), the main drawback (usually ignored) is that the truncation error does not necessarily decrease with the expansion order, that is, high order calculations do not always result in better estimations of the property of interest. In the case of entropy estimations, for instance, if high order correlation effects are significant, there is no guarantee that a second order estimation will improve the sum of marginal entropies (even for infinite sampling).

Herein, we will benefit from the generality of Eq.(2) in order to derive a general, order independent, but cutoff dependent expression. The resulting approach does not contain explicit n -order terms, but implicitly accounts for maximum order effects within a given cutoff, thereby removing the awkward order dependency and the costly combinatorial explosion of the original

1
2
3 expansion. We will also analyze the problem of false correlation in entropy calculations and
4
5 devise a strategy for selecting the best cutoff for a given amount of sampling. Selecting the best
6
7 cutoff is a one of the key points of the present work because, as far as we know, it is
8
9 unprecedented in statistical entropy estimations from Molecular trajectories.
10
11

12
13
14 The applicability of our approach will be demonstrated by computing the conformational entropy
15
16 ($S_{conform}$) of a 45-residue peptide molecule after having characterized its conformational states
17
18 during a classical MD simulation by means of the discretization of the time evolution of the
19
20 internal rotations about single bonds.¹³⁻¹⁴ This entropy contribution arises from the decomposition
21
22 of the total entropy (S_{tot}) of a single molecule (excluding translation and rotation) into a
23
24 vibrational (\bar{S}_{vib}) and a pure conformational contribution ($S_{conform}$), $S_{tot} = \bar{S}_{vib} + S_{conform}$, as
25
26 originally proposed by Karplus and coworkers.¹⁴⁻¹⁷ Finally, we will follow our recently proposed
27
28 computational procedure within the context of Molecular Dynamics simulations¹³ to combine the
29
30 $S_{conform}$ values together with the translational, rotational and the mean vibrational contributions,
31
32 computed through the rigid-rotor harmonic-oscillator (RRHO) approach, in order to estimate
33
34 absolute entropies. We will show that the resulting entropies outperform the broadly used quasi-
35
36 harmonic entropies¹⁸ in terms of both the resulting upper bound entropy value and its
37
38 convergence properties with respect to the simulation time.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

THEORY: *Multibody Local Approximation*

Let us first define

$$\mathcal{C}(R) = \left\{ \mathcal{I} \subset \mathcal{A} \mid \max_{A_i, A_j \in \mathcal{I}} \{d(A_i, A_j)\} < R \right\}$$

as the set composed by all subsets \mathcal{I} of $\mathcal{A} = \{A_1, \dots, A_M\}$ in which a generalized distance

$d(A_i, A_j)$ between any pair of elements A_i and A_j belonging to \mathcal{I} is lower than a predefined cutoff

R . If the inner sum in Eq.(3) is restricted to all the sets \mathcal{I} included in $\mathcal{C}(R)$, then we obtain an R -

dependent n -order MIE approximation to the total entropy:

$$S_R^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{I} \in \mathcal{C}(R) \\ |\mathcal{I}|=k}} I_k(\mathcal{I}). \quad (5)$$

In this way, for a given cardinality k , not all the possible $\binom{M}{k}$ combinations are evaluated, but

only those fulfilling the R -cutoff criterion. Note that the approximation preserves the entropy

invariance under any $A_i \leftrightarrow A_j$ permutation since the number of elements and composition of

$\mathcal{C}(R)$ are independent of how the elements of \mathcal{A} were labeled. Nevertheless, it turns out that

using a reasonable cutoff (e.g., 8-10 Å for conformational entropy calculations), the evaluation of

the *localized* $S_R^{(n)}$ expression is still computationally very expensive at high orders (see SI for

further details).

1
2
3
4 To overcome the practical limitations of $S_R^{(n)}$ we assume first, and without loss of generality,
5
6 that every set $\mathcal{I} = \{A_i, A_j, \dots\}$ included in Eq.(5) is an ordered set (*i.e.*, $i < j < \dots$) for which, by
7
8 definition, the distance between every pair of elements is lower than R . For convenience, we
9
10 propose to relax this condition by requiring only that the distance $d(A_i, A_k)$ between the *first*
11
12 element of \mathcal{I} , which must be chosen to have the lower index i , and any other element A_k of the set
13
14 is less than R . By doing so, it is clear that we will include some *additional* terms not included
15
16 in $\mathcal{C}(R)$, although as we shall see, as a consequence the resulting MIE expression will collapse to
17
18 a much simpler form for maximum order.
19
20
21
22
23
24
25
26
27

28 To this end, we define a *list* \mathcal{L}_i associated to each element A_i of \mathcal{A} as the set
29
30
31 $\mathcal{L}_i = \{A_j \in \mathcal{A} \mid j \geq i, d(A_i, A_j) < R\}$, where each list \mathcal{L}_i comprises the element A_i and its
32
33 neighbors A_j within the predefined cutoff R that lie above A_i in the ordered set \mathcal{A} . Note that the
34
35 upper bound distance between two arbitrary elements A_j and A_l of \mathcal{L}_i is $2R$ because of the triangle
36
37 inequality of the metric space we are working on. Subsequently, in order to approximate Eq.(5),
38
39 we write the following n -order entropy expression in which the sum over the mutual information
40
41 terms is now performed by summing first over an ordered lists $\{\mathcal{L}_1, \dots, \mathcal{L}_M\}$ and then over all the
42
43 ordered subsets \mathcal{J} included in $\mathcal{L}_i - \{A_i\}$:
44
45
46
47
48
49
50
51

$$S_{\mathcal{L}}^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} I_k(A_i, \mathcal{J}), \quad (6)$$

The notation employed in this expression, $S_{\mathcal{L}}^{(n)}(\mathcal{A})$, was chosen to remark that this entropy estimation depends directly on the constructed *lists* for each element A_i rather than on the R cutoff. Similarly, $\{A_i, \mathcal{J}\} = \{A_i, J_1, \dots, J_{k-1}\}$ stands for all the possible subsets that meet our less restrictive requirement, because by construction $d(A_i, J_m) < R$ for every $m \in \{1, \dots, k-1\}$. Finally, $I_k(A_i, \mathcal{J})$ is the mutual information shared among the k elements of $\{A_i, J_1, \dots, J_{k-1}\}$. Notice that $|\mathcal{J}| = 0 \Leftrightarrow \mathcal{J} = \emptyset$ and that by definition $I_1(A_i) = S(A_i)$.

Expressing the mutual information functions in terms of entropies and carrying out various algebraic manipulations, we can transform Eq.(6) without any further approximation into (see SI)

$$S_{\mathcal{L}}^{(n)}(\mathcal{A}) = \sum_{i=1}^M \sum_{k=1}^n \left[\sum_{j=0}^{n-k} (-1)^j \binom{|\mathcal{L}_i| - k}{j} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}| = k-1}} (S(A_i \cup \mathcal{J}) - S(\mathcal{J})), \quad (7)$$

where $S(A_i \cup \mathcal{J})$ represents the joint entropy of the set $\{A_i, J_1, \dots, J_{k-1}\}$, and by definition, $S(\emptyset) = 0$. It must be emphasized that $S_{\mathcal{L}}^{(n)}$ is not strictly invariant under any $A_i \leftrightarrow A_j$ permutation due to the *additional* terms included in $S_{\mathcal{L}}^{(n)}$ that belong to the $\mathcal{C}(2R) - \mathcal{C}(R)$ set and that are not included in $S_R^{(n)}$. Nevertheless, we will see that, the numerical inaccuracies arising from the loss of $A_i \leftrightarrow A_j$ invariance are perfectly assumable. Furthermore, since the number of the additional terms depends on the elements ordering, it can be reduced by properly rearranging the elements of \mathcal{A} , thereby minimizing the numerical effects of the permutation inconsistency in $S_{\mathcal{L}}^{(n)}$ (see SI).

To finally derive the expression of the Multibody Local Approximation, we realize that $S_{\mathcal{L}}^{(n)}$ (Eq.(7)) can be computed so that for each list \mathcal{L}_i , the value of the expansion order n can be selected adaptively by taking its maximum possible value, that is, $n = |\mathcal{L}_i|$. This choice together with the following combinatorial identity, $\sum_{j=0}^m (-1)^j \binom{m}{j} = 0, \forall m \in \{1, 2, \dots\}$,¹⁹ allow us to obtain a largely simplified entropy expression. When $n = |\mathcal{L}_i|$ it turns out that the sum between squared brackets in Eq.(7) is equal to zero unless $k = |\mathcal{L}_i|$. Also, since there is only one subset $\mathcal{J} \subset (\mathcal{L}_i - \{A_i\})$ with cardinality $|\mathcal{L}_i| - 1$, which is $\mathcal{L}_i - \{A_i\}$ itself, then the result of summing over all $k \leq |\mathcal{L}_i|$ is

$$S_{\mathcal{L}}(\mathcal{A}) = \sum_{i=1}^M [S(\mathcal{L}_i) - S(\mathcal{L}_i - \{A_i\})], \quad (8)$$

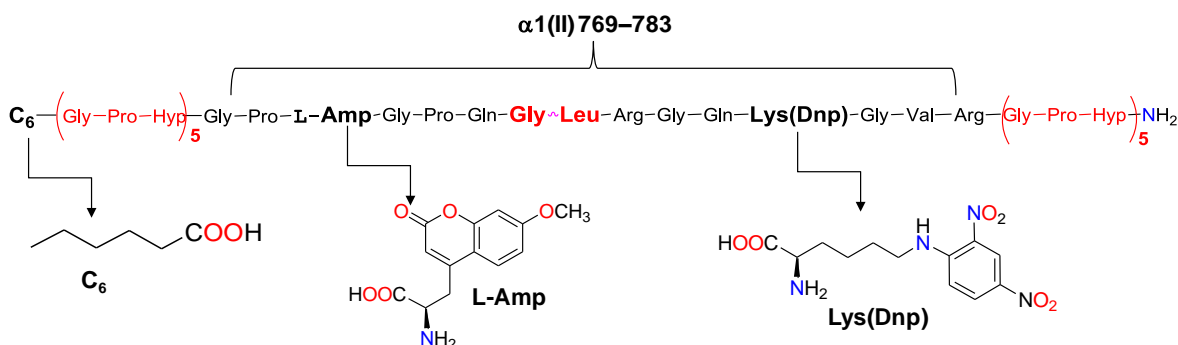
where the resulting entropy estimation $S_{\mathcal{L}}$ corresponds to the Multibody Local Approximation (MLA) for the calculation of $S(\mathcal{A})$, which becomes effectively independent of any expansion order n . MLA evaluates, without computing them explicitly, all-order correlation effects within the given cutoff. As shown in the Supporting Information, it can be easily verified that the MLA expression, Eq.(8), is exact for infinite cutoff ($R=\infty$) likewise the MIE expression, Eq.(2), is exact for maximum order.

At this point it may be convenient to introduce a new definition. We shall say a property S is *local* in the set \mathcal{A} , if $S(\mathcal{A})$ tends to the sum $\sum_{A_i \in \mathcal{A}} S(A_i)$ as long as we increase the generalized distance d among the system elements $\{A_i\}$. MLA entropies, for instance, are local properties

since $S_{\mathcal{L}}(\mathcal{A})$ tends to the sum of marginal entropies whenever that the A_i elements are effectively “decoupled” to each other either by doing $R=0$ or by increasing the distances $d(A_i, A_j)$ between the system components. Fortunately, most of the properties we could be interested in are local. Finally, it can also be noteworthy that the MLA expression in Eq. (8) is similar to the approximations used by different thermochemically-based protocols aimed at determining the total energy of a large system as a combination of fragment energies, which can be considered as truncated forms of the more general multibody expansion approach.²⁰

MLA Test Calculations of Conformational Entropies

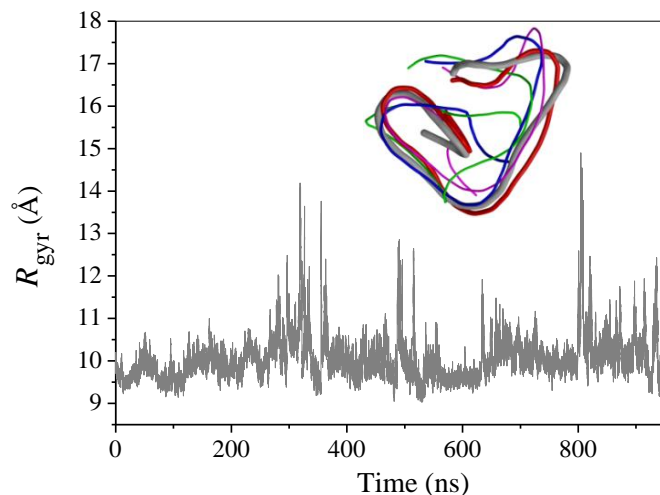
Clearly, the derivation of the MLA expression raises several concerns about its accuracy and computational performance with respect to the order-dependent local approximation $S_R^{(n)}$, the influence of the R cutoff, etc. All these issues were critically analyzed in this work by carrying out conformational entropy calculations of a relatively large biomolecule, which is a challenging problem in entropy calculations that has received much attention due to its importance for understanding the folding and/or association of biomolecular systems.²¹⁻²³



Scheme 1

1
2
3
4
5
6 More specifically, we computed the conformational entropy of the monomeric state of a synthetic
7
8 fluorogenic peptide termed as **fTHP-5** (see Scheme 1), which is a relatively flexible 45-residue
9
10 long synthetic peptide that mimics a segment of the α 1-chain of type II human collagen.²⁴⁻²⁵ The
11
12 1.0 μ s MD trajectory of the **fTHP-5** peptide in explicit solvent, which was started from the initial
13
14 structure favored by the LMOD algorithm, populates conformational regions of the solute
15
16 molecule characterized by a radius of gyration (R_{gyr}) of 10.0 ± 0.5 Å. Although the **fTHP-5**
17
18 backbone chain is rich in *rigid* imidic acids (Pro and Hyp), a significant structural variability was
19
20 observed during the MD simulation. In fact, relatively ample conformational changes passing
21
22 through more extended structures occur during the MD simulation as observed in the time
23
24 evolution of R_{gyr} , which shows frequent peaks having values around 12-13 Å (see Figure 1).
25
26
27 Similarly, clustering analyses also confirm that, although the two ends of **fTHP-5** are rich in *rigid*
28
29 imidic acids, it still exhibits a relatively large dynamical flexibility through its backbone motions
30
31 and, therefore, it is conceivable that conformational entropy could be large enough to play a
32
33 significant role in the global stability of the **fTHP-5** to MMPs.
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1. Time evolution of the fTHP-5 radius of gyration (in Å) and superposition of the most populated representative structures derived from clustering analyses for the fTHP-5 monomer. 500 snapshots were clustered using the MMTSB-tools. The mutual similarity algorithm was employed by selecting a fixed cluster radius (90 degrees) and considering only the backbone torsion angles. The structure in each cluster with the lowest deviation is taken as the cluster representative. Thickness of the models corresponds to the number of snapshots represented by each model.

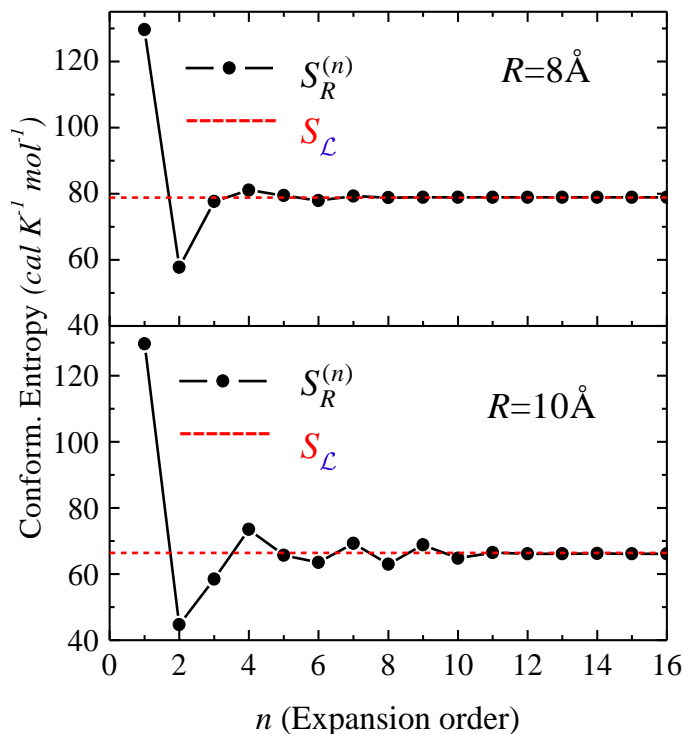


Comparison between $S_R^{(n)}$ and $S_{\mathcal{L}}$ conformational entropies

Once the conformational states of the fTHP-5 peptide along the MD trajectory are characterized following as explained in Computational Methods (see below), we can study the convergence of the cutoff and order-dependent entropy $S_R^{(n)}$ with respect to the expansion order n to find out whether or not the MLA entropy $S_{\mathcal{L}}$ and $S_R^{(n)}$ give comparable results. Given that the

1
2
3 computation of $S_R^{(n)}$ at high orders is very expensive, we used a small fraction of the global data
4
5
6 for these test calculations (2000 MD snapshots). Figure 2 plots the value of $S_R^{(n)}$ versus the
7
8 expansion order (from $n=1$ to 16) for two different cutoff values ($R=8.0$ and 10.0 Å) while the
9
10 corresponding $S_{\mathcal{L}}$ values, which are order independent, are shown as zero-slope lines. It is clear
11
12 that $S_R^{(n)}$ converges to the MLA value although convergence is slower for the larger cutoff value
13
14
15 in consonance with expectations. It must be emphasized, however, that obtaining converged
16
17 estimations of $S_R^{(n)}$ requires the evaluation of the entropy of millions of terms appearing in the
18
19 multibody expansion (*e.g.*, $\sim 10^8$ terms for $R=10$ Å and $n=12$) whereas the MLA approximation
20
21 $S_{\mathcal{L}}$ gives very similar entropy estimations at a much reduced computational cost. More precisely,
22
23 our $S_{\mathcal{L}}$ test calculations resulted in a $\sim 10^4$ CPU time speedup with respect to that consumed by
24
25 the $S_R^{(n)}$ calculations, which in addition needed a larger volume of rapid access memory (see SI
26
27 for further details). In terms of accuracy, the observed differences between the $S_{\mathcal{L}}$ and $S_R^{(n=16)}$
28
29 values are 0.07 and 0.24 *cal/(mol K)* for $R=8$ Å and 10Å, respectively. Hence, the small numerical
30
31 errors of the $S_{\mathcal{L}}$ values, which are due to the loss of invariance of $S_{\mathcal{L}}$ leading to the inclusion of
32
33 additional terms, could be considered as negligible for many real case applications. Therefore,
34
35 these test calculations suggest that the MLA expression could be a relatively cheap alternative for
36
37 computationally intensive entropy calculations with samplings of millions of configurations.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 2: Plots of $S_R^{(n)}$ versus expansion order for the conformational entropy of the f-THP5 peptide using two cutoff values ($R=8$ and 10 \AA). Comparison with the limit value predicted by MLA ($S_{\mathcal{L}}$) is also shown. Only 2000 MD snapshots were used to speed up the calculations.



Cutoff selection and correlation correction

The MLA expression replaces the uncertainty in the empirical decision of which order truncation is more appropriated by another critical question: Which is the best cutoff R for estimating the conformational entropy from a given amount of MD sampling? It might appear that the larger cutoff (if computationally affordable), the better entropy estimation, but this is far from being generally true. In principle, a larger R value should account for more correlation effects among

1
2
3 the system degrees of freedom, but it also means to increase the size of the sample space and,
4
5 consequently, the corresponding *bias* of the entropy estimation, simply because no unbiased
6
7 estimator of entropy exists.²⁶ Therefore, for every particular problem, there must be a cutoff value
8
9 above which the MLA entropy estimation becomes unreliable due to the large size of the entropy
10
11 bias. Unlike traditional methods for entropy estimation like the classical Miller-Madow bias
12
13 correction,²⁷ Bayesian alternatives²⁸⁻²⁹ or the relatively recent Chao-Shen estimator;³⁰ herein we
14
15 propose a more empirical strategy based on the outcomes of our *statistical experiment* (i.e., the
16
17 MD simulation).
18
19
20
21
22
23

24
25 To find an optimal cutoff value for the MLA conformational entropy calculations, we first need
26
27 reasonably converged marginal entropies of all the rotatable bonds. Normally only 1-3
28
29 conformational states are accessible to each torsion angle A_i and the underlying probability mass
30
31 functions $P(A_i)$ should be quite converged after having included about 10^6 configurations in the
32
33 calculations (e.g., for fTHP-5 the convergence gradient of $-T \sum_i^M S(A_i)$ was found to be
34
35 $6 \times 10^{-3} \text{ kcal mol}^{-1} \text{ ns}^{-1}$). In this scenario, it is reasonable to assume that all the entropy bias is
36
37 entirely due to *false* correlation effects, and then we can rewrite Eq.(3) in order to express the
38
39 *exact* entropy of a given system \mathcal{A} for a finite amount of sampling as:
40
41
42
43
44
45

$$S(\mathcal{A}) = \sum_{A_i \in \mathcal{A}} S(A_i) + \sum_{k=2}^M (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=k}} \left\{ E[\tilde{I}_k(\mathcal{J})] - B[\tilde{I}_k(\mathcal{J})] \right\}, \quad (9)$$

46
47 where $S(A_i)$ are the “exact” values of the marginal entropies, $E[\tilde{I}_k(\mathcal{J})]$ is the expected
48
49 value of the *sample* mutual information $\tilde{I}_k(\mathcal{J})$ shared among the components of \mathcal{J} . On the
50
51
52
53
54
55
56
57
58
59
60

other hand, $B[\tilde{I}_k(\mathcal{J})]$ is the bias of $\tilde{I}_k(\mathcal{J})$ due to the finiteness of the sample. Notice that we are now (and temporarily) using tildes “ \sim ” to denote *sample* magnitudes.

Unfortunately, there is no exact way to predict this bias,²⁶ and, therefore, we devised the following strategy that seems particularly suitable for guessing the value of the conformational entropy bias in molecular trajectory simulations. First, we formally define for every subset of system variables \mathcal{J} with k elements a discrete random (in the statistical sense) vector $X = (X_1, \dots, X_k)$ specifying the corresponding conformational state, that is, the conformational state of the i -th torsion included in \mathcal{J} is given by the value of X_i . Each component X_i has its own set of outcomes $\{x_i(t_1), x_i(t_2), \dots\}$ collected along the MD simulation. If the outcomes of at least $k-1$ components of X are *independently* and *randomly* reordered, then the correlation among the components (variables) of X is artificially removed, and at the same time, we obtain the outcomes of a hypothetical random variable $X' = (X'_1, \dots, X'_k)$ with independent components.

As the correlation is destroyed, the exact mutual information function among the X' components, $I_k(X'_1, \dots, X'_k) \equiv I'_k(\mathcal{J})$ must be null and, by the own bias definition,³¹ we have that $B[\tilde{I}'_k(\mathcal{J})] = E[\tilde{I}'_k(\mathcal{J})]$. Now, since the bias of the mutual information is, to first order, equal to the bias of the independent case,³²⁻³³ for a sufficiently long period of simulation, we can approximate $B[\tilde{I}_k(\mathcal{J})]$ to $B[\tilde{I}'_k(\mathcal{J})]$, transforming thus Eq.(9) into

$$\hat{S}(\mathcal{A}) = \sum_{A_i \in \mathcal{A}} S(A_i) + \sum_{k=2}^M (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=k}} E[\tilde{I}_k(\mathcal{J}) - \tilde{I}'_k(\mathcal{J})]. \quad (10)$$

Under the same conditions, the variances of the entropy and the mutual information are vanishing when compared to their bias,²⁶ and the expectation $E[\tilde{I}_k - \tilde{I}'_k]$ in Eq. (10) can be replaced by

$\tilde{I}_k - \tilde{I}'_k$ obtaining the entropy estimator

$$\hat{S}(\mathcal{A}) = \sum_{A_i \in \mathcal{A}} S(A_i) + [\tilde{S}(\mathcal{A}) - \tilde{S}'(\mathcal{A})], \quad (11)$$

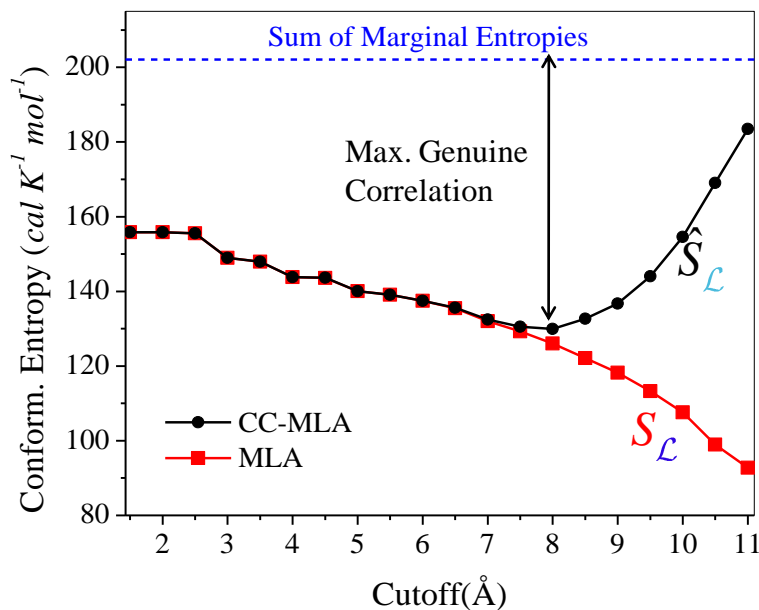
where $\tilde{S}(\mathcal{A})$ is the sample entropy and $\tilde{S}'(\mathcal{A})$ is the sample entropy computed for the M -dimensional random variable X' representing the conformational state of \mathcal{A} after having independently and randomly reordered the outcomes of at least $M-1$ of its components. It is important to emphasize that the transformations made from Eq.(9) would be exact for an infinite sampling and therefore $\hat{S}(\mathcal{A})$ must be asymptotically unbiased. This new entropy estimator can be used directly with the corresponding multibody local approximations to obtain the correlation corrected MLA (CC-MLA) entropy estimation (Eq.(12)). From now on, all the entropies we will deal with are sample entropies, so the tildes will be omitted hereinafter.

$$\hat{S}_{\mathcal{L}}(\mathcal{A}) = \sum_{A_i \in \mathcal{A}} S(A_i) + [S_{\mathcal{L}}(\mathcal{A}) - S'_{\mathcal{L}}(\mathcal{A})]. \quad (12)$$

Figure 3 shows the behavior of the CC-MLA entropy estimator $\hat{S}_{\mathcal{L}}$ at various cutoff values as compared with that of the uncorrected MLA entropy $S_{\mathcal{L}}$. In these calculations, we employed $\sim 10^6$ MD configurations taken from the 1.0 μ s trajectory of the fTHP-5 peptide molecule. At low values of the cutoff threshold ($R < 8 \text{ \AA}$), the two entropy estimations result in very similar values that progressively decrease with R because of the inclusion of long-range correlation effects. In

1
2
3 other words, the amount of sampling ($\sim 1.0 \mu\text{s}$) turned out to be enough for capturing all the
4
5 genuine correlation among the torsion angles whose mean distance during the MD trajectory is
6
7 below 8 \AA . However, as R increases, we see a turning point in the behavior of the CC-MLA
8
9 entropy warning us that the dimension of the statistical sample space associated to the \mathcal{L}_i and
10
11 $\mathcal{L}_i - \{A_i\}$ sets is becoming too large in comparison with the fixed amount of MD sampling, what
12
13 leads to the emergence of *false* correlation involving relatively far separated torsion angles. As a
14
15 consequence, the uncorrected MLA entropy $S_{\mathcal{L}}$ that includes both true and false correlation
16
17 effects still decreases monotonically with R whereas the trend in the CC-MLA entropy estimator,
18
19 which presumably removes all the false correlation, is reversed, thereby giving larger (more
20
21 positive) entropies. Eventually, both entropy estimations (MLA and CC-MLA) become clearly
22
23 unreliable at large values of R where the dimensionality of the sample space is overwhelming.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 3. Comparison between MLA and CC-MLA. Note that the plot intentionally begins in $cutoff=1.5\text{\AA}$ because, in our distance definition, the entropy below this value is the sum of marginal entropies for both estimators.



To finally answer the question of what is the best cutoff for the MLA calculations, we realize that, for a given cutoff, the value of the $S_R^{(n=M)}$ entropy would give a rigorous upper bound to the exact entropy provided that no entropy bias exists. Since the MLA is a good approximation of $S_R^{(n=M)}$, we would also expect it to be an upper bound of the exact value. By fulfilling this ideal condition, the MLA calculations with larger cutoffs would capture more correlation effects that contribute negatively to the total entropy until converging to the exact entropy. However, the inclusion of false correlation invalidates the MLA entropy S_L as an upper bound to the exact

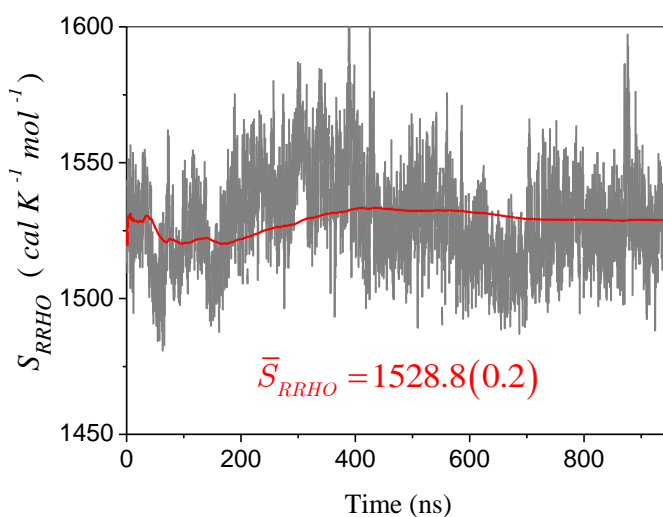
1
2
3
4 entropy. In contrast the CC-MLA estimator $\hat{S}_{\mathcal{L}}$ is constructed to be essentially free from the
5
6
7 entropy bias due to the false correlation, and thereby $\hat{S}_{\mathcal{L}}$ can be considered as an upper bound to
8
9
10 the true entropy at whatever cutoff (although useless if R is too large and the MD sampling is too
11
12 small). Therefore, the minimum value of $\hat{S}_{\mathcal{L}}$ as a function of R allow us to identify the best
13
14 cutoff for the CC-MLA entropy calculations as it provides the lowest upper bound to the exact
15
16 entropy ($R \sim 8 \text{ \AA}$ in the case of the fTHP-5 peptide, see Figure 3). Under our approach, the
17
18 difference between $\min\{\hat{S}_{\mathcal{L}}\}$ and the sum of marginal entropies represents the maximum
19
20 amount of genuine correlation that can be retrieved from the available MD sampling.
21
22
23
24
25
26
27

28 ***Combining CC-MLA and RRHO entropies***

29
30 As mentioned in the Introduction, we have recently proposed¹³ a protocol for entropy calculations
31
32 of single molecules that combines the mean values of the RRHO entropic contributions with the
33
34 conformational entropy derived from the converged probability mass functions of the discretized
35
36 torsion angles and using the classical MIF expansion (Eq. (3)). This approach has been shown to
37
38 capture high order correlation effects and simultaneously yield converged conformational
39
40 entropies of small and medium-sized molecules within a reasonable simulation time. However,
41
42 for larger molecules, we find serious convergence problems with respect to the MIE order that is
43
44 required to capture correlation effects in the conformational entropy. For instance, classical MIE
45
46 estimations of the conformational entropy of fTHP-5 are highly unstable even at second order
47
48 due to the problem of long-range false correlation and the fact that truncation errors do not
49
50 decrease monotonically with increasing order expansion (see SI). In contrast, the CC-MLA
51
52 approximation with $R=8 \text{ \AA}$ allowed us to obtain optimized and well converged values of
53
54
55
56
57
58
59
60

conformational entropies that can be subsequently combined with the mean values of the vibrational entropy computed by carrying out normal mode calculations under the RRHO approximation on a set of representative MD snapshots (see Figure 4).

Figure 4. Time evolution of the RRHO entropy of fTHP-5 computed by normal mode calculations on 5000 snapshots.



On the other hand, the time evolution of the MLA and CC-MLA methods for the optimal cutoff (8 Å) is shown in Figure 5. As can be seen, in spite of the apparent time convergence of the MLA entropy, the entropy bias is still present as suggests the difference between both approximations.

Finally, Figure 6 shows the convergence of the absolute entropy ($S \approx \bar{S}^{RRHO} + \hat{S}_{\mathcal{L}}^{conform}$) obtained by adding the RRHO entropy to the CC-MLA conformational term at the optimal cutoff (8Å). The RRHO term, which was computed by normal mode analyses using a solvent continuum model,³⁴ accounts for the majority of S (95%) and converges quite rapidly. However, Figure 5 reveals that S only reaches a stable plateau at $\sim 1611 \text{ cal } K^{-1} \text{ mol}^{-1}$ once that the $\hat{S}_{\mathcal{L}}^{conform}$ (CC-

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

MLA) term is converged after ~600 ns. Note that the limiting value of the RRHO & CC-MLA entropy S constitutes an upper bound to the actual entropy value because correlation effects between vibrational motions and conformational changes are neglected by the above entropy partitioning¹³ and long range correlations in $\hat{S}_{\mathcal{L}}^{conform}$ are likewise neglected by the use of the cutoff.

Figure 5. Time evolution of the conformational entropy using the optimal cutoff (8Å) using MLA and its correlation corrected version CC-MLA.

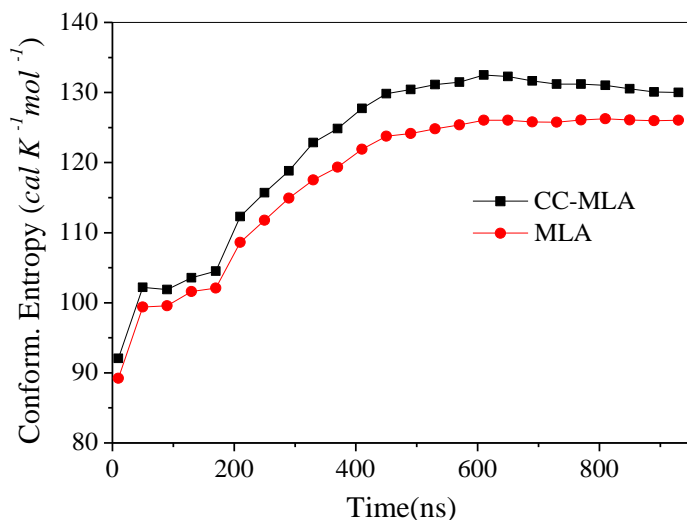
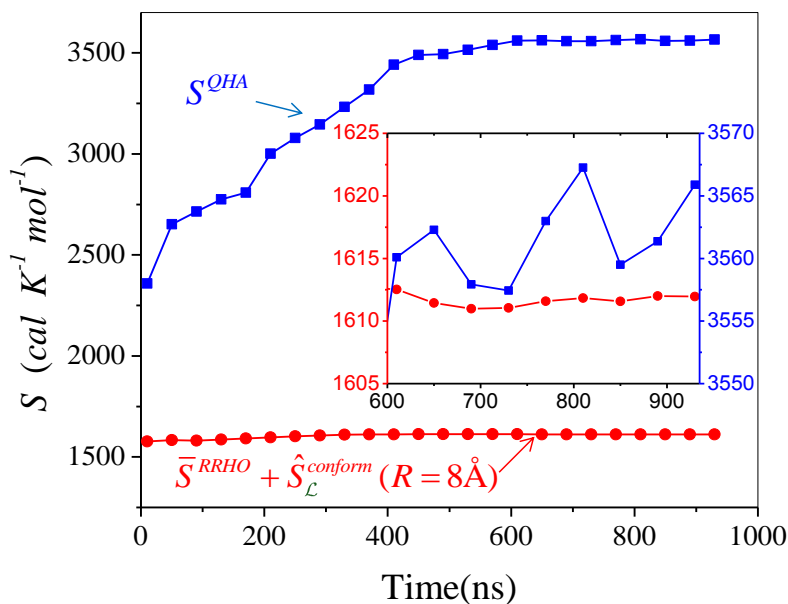


Figure 6. Comparison between S^{QHA} and the partition scheme $\bar{S}^{RRHO} + \hat{S}_{\mathcal{L}}^{conform}$. The inset zooms in the last 0.33 μs showing the different convergence behavior of S^{QHA} and $\bar{S}^{RRHO} + \hat{S}_{\mathcal{L}}^{conform}$.



For comparative purposes, we also computed the quasi-harmonic approximation (QHA) entropy¹⁸ of f-THP5, resulting in a very large entropy value (~ 3560 cal/mol K), which is 1954 cal K^{-1} mol $^{-1}$ above that provided by the RRHO & CC-MLA calculation (580 kcal mol $^{-1}$ at 298 K in terms of free energy). Thus, the well-known limitations of the QHA method³⁵⁻³⁶ (e.g., neglecting supralinear correlations, approximating multimodal distributions to unimodal one, etc.) ultimately result in a non-systematic and much larger overestimation of S than the RRHO & CC-MLA approach. Furthermore, the QHA calculations also exhibit worse convergence properties (see the inset in Figure 6).

SUMMARY

Computation of conformational entropies from MD simulations of large molecules using the classical MIE expansion presents a serious convergence problem with reference to the expansion order that is required to capture correlation effects. In this work we have shown that this problem can be substantially mitigated by reformulating the general n -order dependent MIE into a localized form (MLA) that depends on a single cutoff parameter R . Our validation calculations indicate that MLA is computationally very efficient at the cost of introducing a small numerical error due to the loss of permutation invariance of the MLA expression. The best cutoff R for a given system can be selected systematically by means of a correlation corrected entropy estimator (CC-MLA) that extracts the maximum amount of genuine correlation from the available MD sampling. This idea could be extended to any statistical entropy estimation defining an appropriate metric space (e.g., using mutual information based distances³⁷). The combination of the CC-MLA conformational entropy with mean values of RRHO entropic terms leads to reasonably converged upper bounds of absolute entropy for relatively large molecules that are much lower than those provided by the popular QHA method. Finally, we note that the MLA expression (Eq.(8)) could be readily applicable to other local properties like potential energy, electronic density, chemical shifts, etc.; given that for these properties, in principle, the MLA truncation error can be reduced arbitrarily by increasing the cutoff. However, in the case of entropy estimations, besides using larger cutoff values for including as much correlation as possible, we also need to include statistical corrections for removing the ubiquitous entropy bias.

COMPUTATIONAL METHODS

MD simulation

Our model peptide fTHP-5 was subjected to a 1.0 μ s classical MD simulations in explicit solvent from which $\sim 10^6$ configurations were extracted for the subsequent entropy calculations. The AMBER03 force field was used to represent the standard amino acids while the required parameters for the hydroxyproline residues (Hyp) were taken from those derived in a previous work.¹⁴ For the rest of non-standard residues in fTHP-5 (see Scheme 1): the terminal hexanoic acid (C6), the fluorescent L-Amp residue and the quencher residue Lys(Dnp), new parameters were obtained following the prescriptions for parameter derivation as described in the AMBER03 protocol.³⁸ Thus, initial geometries for these residues were capped with terminal Ace and Nme groups (C6-Nme, Ace-Lamp-Nme and Ace-Lys(Dnp)-Nme) and fully optimized at the HF/6-31G(d) level. Subsequently, single-point B3LYP/cc-pVTZ calculations were carried out using the IEF-PCM continuum solvent model³⁹ to mimic an organic solvent environment ($\epsilon = 4.0$) as implemented in the Gaussian03 program.⁴⁰ From the B3LYP/cc-pVTZ electrostatic potential, atomic partial charges were obtained using the RESP methodology.⁴¹ During the RESP fitting procedure, we imposed the AMBER03 charges for the Ace and Nme residues. The bond, angle, dihedral and Lennard-Jones parameters were available from the AMBER03 database.

Starting coordinates were obtained from conformational search calculations using the LMOD program linked to the Amber10 package.⁴²⁻⁴³ In the LMOD calculations, we employed the augmented AMBER03 force field coupled with the Hawkins-Cramer-Truhlar pairwise Generalized-Born (GB) model.⁴⁴ The lowest energy LMOD structure of fTHP-5 was then

1
2
3 surrounded by a periodic truncated octahedral box of TIP3P water molecules that extended ~ 12 Å
4 from the protein atoms (~ 7600 water molecules plus two Cl^- counterions). The solvent molecules
5 were initially relaxed by means of energy minimizations and 50 ps of MD. Subsequently, the full
6 system was minimized and heated gradually to 300 K during 50 ps of MD. During the MD
7 simulation, the system remains coupled to a thermal and a hydrostatic bath at $T = 300$ K and $P =$
8 1.0 atm, the time step of integration was 2.0, the SHAKE procedure on the X-H bonds was
9 applied, and the PME approach was used for non-bonded interactions. A 1.0 μs trajectory was
10 computed and coordinates were saved for analysis every ps. The first 60 ps of the trajectory were
11 considered as an equilibration period and therefore were not used for entropy calculations. All the
12 energy minimization and MD calculations were carried out using the PMEMD program included
13 in the Amber10 suite.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32 *Conformational entropy calculations*

33
34 In previous work,¹³ we have described in full detail our procedure for discretizing the probability
35 density functions of those torsion angles that are commonly used to define the molecular
36 conformational state, prior to the conformational entropy calculations. To this end, we collect the
37 time series with the values of the rotatable torsion angles along the MD simulation. For each
38 torsion angle θ , we obtain an analytic representation of the underlying probability density
39 function using the von Mises kernel estimator. After having located all the minima and maxima
40 in the configurational space of θ defined by the $[0, 2\pi)$ interval, we obtain a series of m non-
41 overlapping intervals that, in turn, define the different conformational states accessible to θ . In
42 this way, the initial time series containing N data points, $\{\theta_1, \dots, \theta_N\}$ is transformed into a set of N
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 integer numbers $\{x_1, \dots, x_N\}$ labeling the conformational states populated by the torsion angle.
4
5
6 Thus, the continuous variable θ characteristic of the torsion angle becomes a discrete random
7
8 variable X , whose probability mass function, $P(X)$, can be estimated by the frequency of the
9
10 outcomes $\{x_1, \dots, x_N\}$. Analogously, the conformational state of a set of M torsional angles can be
11
12 described by a M -dimensional random vector (X_1, \dots, X_M) , where X_i specifies the
13
14 conformational state of the i -torsion, and M is the number of torsion angles of our system
15
16 $\mathcal{A} = \{A_1, \dots, A_M\}$. This transformation was performed by using the CENCALC software,⁴⁵ which
17
18 uses both trajectory coordinates and topology information in order to characterize the
19
20 conformational states of the molecule of interest by discretizing automatically the time evolution
21
22 of internal rotations. CENCALC also implements various reformulations of the MIE methods
23
24 including MLA in order to estimate the conformational entropy of the molecule of interest.
25
26
27
28
29
30
31
32
33

34 *RRHO entropy calculations*

35
36 Normal mode calculations for the fTHP-5 molecule were carried out using the AMBER03 force
37
38 field and using an implicit solvent model for removing the explicit consideration of the solvent
39
40 degrees of freedom.³⁴ Then, 5000 equally-spaced snapshots were extracted from the 1.0 μ s
41
42 trajectory and post-processed through the removal of all solvent and counterions molecules. Prior
43
44 to the normal mode calculations, the geometries of the system described by the AMBER03 force
45
46 field were minimized until the root-mean-squared of the elements in the gradient vector was less
47
48 than 10^{-5} kcal/(mol Å). These minimizations and the subsequent normal mode calculations³⁴
49
50 were carried out using the HCT-GB solvent model. Finally, the Rigid-Rotor Harmonic-Oscillator
51
52 (RRHO) entropic contributions were averaged over the 5000 snapshots.(see Figure 4) All the
53
54
55
56
57
58
59
60

1
2
3 required energy minimizations and normal mode calculations were performed using the NAB
4
5 package.⁴⁶
6
7
8

9 10 **SUPPORTING INFORMATION**

11
12 Propositions (1-4) and their corresponding mathematical proofs. Strategy followed in the
13
14 reordering the elements of \mathcal{A} in order to minimize the number of what we call additional terms.
15
16

17
18 Table S1 showing the dependence of $S_R^{(n)}$ and its computational cost on the order and cutoff. A
19
20 critical discussion on the applicability of the classical MIE approach; the discussion includes
21
22 Figure S1 and Table S2 that show a poor order convergence of the MIE results in our system.
23
24
25

26
27 This material is available free of charge via the Internet at <http://pubs.acs.org>.
28
29
30
31

32 **ACKNOWLEDGMENT**

33
34
35
36 This research was supported by the following grants: FI-CyT (Asturias, Spain) IB05-076 and
37
38 MEC (Spain) CTQ2007-63266. The authors thank Dr. Angel Martín Pendás for his careful
39
40 reading of the manuscript and his suggestions and Drs. Haydee Valdés and Natalia Díaz for
41
42 having derived new molecular mechanics parameters and started the MD simulation of the fTHP-
43
44 5 peptide.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

- (1) Drautz, R.; Fähnle, M.; Sanchez, J. M. General relations between many-body potentials and cluster expansions in multicomponent systems. *J. Phys. Condens. Matter* **2004**, *16*, 3843-3852.
- (2) Sundararaghavan, V.; Zabaras, N. Weighted multibody expansions for computing stable structures of multiatom systems. *Phys. Rev. B* **2008**, *77*, 064101-064110.
- (3) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29*, 1605–1614.
- (4) Matsuda, H. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E* **2000**, *62*, 3098-3102.
- (5) Babu, K.; Gadre, S. R. Ab Initio Quality One-Electron Properties of Large Molecules: Development and Testing of Molecular Tailoring Approach. *J. Comput. Chem.* **2003**, *24*, 484-495.
- (6) Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Correlation Energy, with Applications to the Calculation of Accurate Second-Order Miller-Plesset Perturbation Theory Energies for Large Water Clusters *J. Chem. Theory Comput.* **2007**, *3*, 1342 - 1348.
- (7) Huang, L.; Massa, L.; Karle, J. The kernel energy method of quantum mechanical approximation carried to fourth-order terms. *Proc. Natl. Acac. Sci. USA* **2008**, *105*, 1849–1854.
- (8) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. Fragment molecular orbital method: analytical energy gradients. *Chem. Phys. Lett.* **2001**, *336*, 163-170.
- (9) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. Application of the Electrostatically Embedded Many-Body Expansion to Microsolvation of Ammonia in Water Clusters. *J. Chem. Theory Comput.* **2008**, *4* 683–688.
- (10) Xantheas, S. S. Ab initio studies of cyclic water clusters (H₂O)_n, n=1-6. II. Analysis of many-body interactions. *J. Chem. Phys.* **1994**, *100*, 7523-7534.
- (11) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **2007**, *127*, 024107-024116.
- (12) Dahlke, E. E.; Truhlar, D. G. Assessment of the Pairwise Additive Approximation and Evaluation of Many-Body Terms for Water Clusters. *J. Phys. Chem. B* **2006**, *110*, 10595-10601.
- (13) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2638-2653.
- (14) Suárez, E.; Díaz, N.; Suárez, D. Entropic Control of the Relative Stability of Triple-helical Collagen Peptide Models. *J. Phys. Chem. B* **2008**, *112*, 15248–15255.
- (15) Chang, C. A.; Chen, C.; Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acac. Sci. USA* **2007**, *104*, 1534–1539.
- (16) Karplus, M.; Ichiye, T.; Pettit, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083-1085.
- (17) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092-4107.

- 1
2
3 (18) Andricioaei, I.; Karplus, M. On the calculation of entropy from covariance
4 matrices of the atomic fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289-6292.
- 5 (19) Comtet, L. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*;
6 Reidel Publishing Company: Holland, 1974.
- 7 (20) Suárez, E.; Díaz, N.; Suárez, D. Thermochemical Fragment Energy Method for
8 Biomolecules: Application to a Collagen Model Peptide. *J. Chem. Theory Comput.* **2009**, *5*,
9 1667–1679.
- 10 (21) Choa, S. S.; Levya, Y.; Wolynesa, P. G. Quantitative criteria for native energetic
11 heterogeneity influences in the prediction of protein folding kinetics. *Proc. Natl. Acad. Sci. USA*
12 **2009**, *106*, 434-439.
- 13 (22) Bachmann, A.; Kiefhaber, T.; Boudko, S.; Engel, J.; Bächinger, H. P. Collagen
14 triple-helix formation in all-trans chains proceeds by a nucleation growth mechanism with a
15 purely entropic barrier. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13897-13902.
- 16 (23) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational entropy
17 changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J. Biomol.*
18 *NMR* **2009**, *45*, 157-169.
- 19 (24) Lauer-Fields, J. L.; Tuzinski, K. A.; Shimokawa, K.-I.; Nagase, H.; Fields, G. B.
20 Hydrolysis of Triple-helical Collagen Peptide Models by Matrix Metalloproteinases. *J. Biol.*
21 *Chem.* **2000**, *275*, 13282-13290.
- 22 (25) Lauer-Fields, J. L.; Kele, P.; Sui, G.; Nagase, H.; Leblanc, R. M.; Fields, G. B.
23 Analysis of matrix metalloproteinase triple-helical peptidase activity with substrates
24 incorporating fluorogenic L- or D-amino acids. *Anal. Biochem.* **2003**, *321*, 105-115.
- 25 (26) Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation*
26 **2003**, *15*, 1191-1253.
- 27 (27) Miller, G. In *Information Theory in Psychology: Problems and Methods*; Quastler,
28 H., Ed.; The Free Press: 1955, p 95-100.
- 29 (28) Holste, D.; Große, I.; Herzog, H. Bayes' estimators of generalized entropies. *J.*
30 *Phys. A: Math. Gen.* **1998**, *31*, 2551-2566.
- 31 (29) Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos*
32 **1996**, *6*, 414-427.
- 33 (30) Chao, A.; Shen, T.-J. Nonparametric estimation of Shannon's index of diversity
34 when there are unseen species. *Environ. Ecol. Stat.* **2003**, *10*, 429-443.
- 35 (31) Rohatgi, V. K.; Saleh, A. K. M. E. *An Introduction to Probability and Statistics*,
36 2001.
- 37 (32) Treves, A.; Panzeri, S. The Upward Bias in Measures of Information Derived from
38 Limited Data Samples. *Neural Computation* **1995**, *7*, 399-407.
- 39 (33) Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different
40 information measures. *Computation in Neural Systems* **1996**, *7*, 87-107.
- 41 (34) Brown, R. A.; Case, D. A. Second Derivatives in Generalized Born Theory. *J.*
42 *Comput. Chem.* **2006**, *27*, 1662–1675.
- 43 (35) Chang, C.; Chen, W.; Gilson, M. K. Evaluating the accuracy of the quasiharmonic
44 approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017-1028.
- 45 (36) Baron, R.; Gunsteren, W. F. v.; Hünenberger, P. H. Estimating the configurational
46 entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the
47 quasi-harmonic approximation. *Trends in Physical Chemistry* **2006**, *11*, 88-122.
- 48
49
50
51
52
53
54
55
56
57
58
59
60

(37) Kraskov, A.; Stögbauer, H.; Andrzejak, R. G.; Grassberger, P. Hierarchical clustering using mutual information *Europhys. Lett.* **2005**, *70*, 278-284.

(38) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *14*, 1999-2012.

(39) Tomasi, J.; Mennucci, B.; Cancès, E. The IEF version of the PCM solvation method: An overview of a new method addressed to study molecular solutes at the QM ab initio level. *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211-226.

(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, J. T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. 2003.

(41) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269-10280.

(42) Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, K. F.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A.; University of California: San Francisco, 2008.

(43) Kolossváry, I.; Guida, W. C. Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011-5019.

(44) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices *J. Am. Chem. Soc.* **1998**, *120*, 9401-9409.

(45) Suárez, E.; Díaz, N.; Méndez, J.; Suárez, D. CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations. *J. Chem. Inform. Model.* (Submitted) **2011**.

(46) Macke, T.; Case, D. A. *Modeling unusual nucleic acid structures.*; American Chemical Society: Washington, DC, 1998.

MultiBody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules

Ernesto Suárez* and Dimas Suárez

E-mail: ernesto@fluor.quimica.uniovi.es

Supporting Information

1. Propositions and Mathematical Proofs

A computational shortcoming of the usual form of the Mutual Information Expansion (MIE) for the total entropy is that the evaluation of the sum $\sum_{\mathcal{I} \subset \mathcal{A} / |\mathcal{I}|=k} I_k(\mathcal{I})$ (see Eq. 3 and 4 in the main text) requires, for each value of the index k , the computation of the entropy over all the subsets with cardinality less or equal to k , including those sets with cardinality less or equal to $k-1$ previously computed in I_{k-1} , and so on. In a previous work¹ we have shown that this large redundancy can be substantially avoided by reformulating the MIE equation. For the sake of completeness, we reproduce herein the formal proof of such transformation.

Proposition 1. *The usual truncation of mutual information expansion for the estimation of the total entropy of a system composed of M ensembles (i.e., subsystems) including correlation effects up to n -order with $n \leq M$ is*

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}), \quad (\text{S.1})$$

where \mathcal{J} runs over all possible subsets of $\mathcal{A} = \{A_1, \dots, A_M\}$ with k elements and the mutual information I_k shared among k ensembles is expressed in terms of the subsystem entropies as

$$I_k(\mathcal{J}) = \sum_{l=1}^k (-1)^{l+1} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=l}} S(\mathcal{I}). \quad (\text{S.2})$$

We can affirm that equation (S.1) is equal to

$$S^{(n)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.3})$$

Proof. To proof the equivalence between (S.1) and (S.3), we must verify that the equality (S.1)=(S.3) holds for $n=1$, because this particular case will be the *seed* for a formal proof of the general case by means of mathematical induction. According to the induction principle, a statement $P(n)$ is true $\forall n \in \mathbb{N}$ if the following two statements hold true:

1. The statement is true for the first element ($n=0$ or $n=1$), in our case: $n=1$.
2. If the statement is true for an arbitrary n then it is true for $(n+1)$

Let us first show that (S.1)=(S.3) is valid for $n=1$. Notice that, from the definition of I_k in (S.2), it follows that for $k=1$ the set $\mathcal{J} = \{J_1\}$ has only one element, which is ultimately one of the i -elements of \mathcal{A} , i.e., $\mathcal{J} = \{A_i\}$. Since $\{A_i\}$ itself, is the only subset of $\{A_i\}$ with cardinality one, then $I_1(A_i) = S(A_i)$. Thus, we have that:

$$S^{(1)}(\mathcal{A}) = \sum_{\substack{\mathcal{J} \subset \{A_1, \dots, A_M\} \\ |\mathcal{J}|=1}} S(\mathcal{J}) = \sum_{A_i \in \{A_1, \dots, A_M\}} S(A_i) = \sum_i S(A_i). \quad (\text{S.4})$$

Next we obtain the same result from (S.3). Note that if $n=1$ in (S.3), then $i=0$ and the (S.3)equation is transformed to

$$S^{(1)}(\mathcal{A}) = \binom{M-1}{0} \sum_{\substack{\mathcal{I} \subset \{A_1, \dots, A_M\} \\ |\mathcal{I}|=1}} S(\mathcal{I}),$$

where by definition $\binom{M-1}{0} = \frac{(M-1)!}{((M-1)-0)!0!} = 1$. Finally, we simplify the notation of the

single sum in the latter expression:

$$S^{(1)}(\mathcal{A}) = \sum_{\substack{\mathcal{I} \subset \{A_1, \dots, A_M\} \\ |\mathcal{I}|=1}} S(\mathcal{I}) = \sum_{A_i \in \{A_1, \dots, A_M\}} S(A_i) = \sum_i S(A_i),$$

obtaining thus the same result as in (S.4). Therefore, the statement is true for $n=1$.

The second step in our formal proof of (S.1)=(S.3) is to verify that, if the statement is true for an arbitrary n , then it is true for $(n+1)$. Thus, particularizing the equation (S.3) for $n+1$:

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^{n+1} \left[\sum_{i=0}^{n+1-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.5})$$

Next we split the outer sum by extracting the last term ($k = n+1$):

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n+1-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}) \quad (\text{S.6})$$

Similarly, we split the middle sum in the first term for $i = n-k+1$:

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} + (-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}) \quad (\text{S.7})$$

Under the hypothesis that equation (S.3) is true for n , the last equation can be rewritten as

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + \sum_{k=1}^n \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) + \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=n+1}} S(\mathcal{I}), \quad (\text{S.8})$$

where the two separate sums can be regrouped, obtaining

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + \sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}). \quad (\text{S.9})$$

Following analogous steps, expression (S.1) can be particularized for $n+1$ as

$$S^{(n+1)}(\mathcal{A}) = \sum_{k=1}^{n+1} (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} I_k(\mathcal{J}), \quad (\text{S.10})$$

and then transformed into

$$S^{(n+1)}(\mathcal{A}) = S^{(n)}(\mathcal{A}) + (-1)^n \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} I_{n+1}(\mathcal{J}). \quad (\text{S.11})$$

By comparing (S.9) and (S.11), it is clear that the proof of the original proposition implies that the following equality must hold true:

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = (-1)^n \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} I_{n+1}(\mathcal{J}). \quad (\text{S.12})$$

Therefore we must prove equality (S.12). First, we express I_{n+1} in the right side of (S.12) in terms of the subsystem entropies (for convenience, the l -index used in definition (S.2) is replaced now by k):

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} \sum_{k=1}^{n+1} (-1)^{n+k+1} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k}} S(\mathcal{I})$$

Reordering the sums in the right side and knowing that $(-1)^{n-k+1} = (-1)^{n+k+1}$:

$$\sum_{k=1}^{n+1} \left[(-1)^{n-k+1} \binom{M-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}) = \sum_{k=1}^{n+1} (-1)^{n-k+1} \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=n+1}} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k}} S(\mathcal{I}).$$

Finally, to prove that this last expression is true, we transform the double sum over the subsets \mathcal{J} and \mathcal{I} in the right side into a single sum. To this end we need to count the number of times a given subset \mathcal{I} appears while summing over \mathcal{J} . Once the k elements of \mathcal{I} are selected, there are $(n+1)-k$ unselected elements of \mathcal{J} . Thus, the number of times a given \mathcal{I} will appear is the number of possibilities for selecting $(n+1)-k$ elements from the rest $M-k$ elements, which is exactly $\binom{M-k}{n-k+1}$. Therefore, expression (11) is true, thereby demonstrating the validity of the

Proposition 1.

Proposition 2. *The total entropy, or any other property or function S , of any system $\mathcal{A} = \{A_1, \dots, A_M\}$ composed by M elements, which can be computed on all possible subsets of \mathcal{A} , can be calculated without any approximation as*

$$S(\mathcal{A}) = \sum_{k=1}^M \sum_{\substack{\mathcal{J} \subset \mathcal{A} \\ |\mathcal{J}|=k}} \sum_{l=1}^k (-1)^{k+l} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=l}} S(\mathcal{I}), \quad (\text{S.13})$$

where \mathcal{J} runs over all possible subsets of \mathcal{A} containing k elements while \mathcal{I} runs over all possible subsets of \mathcal{J} .

Proof. Note that Eq.(S.13) is no other than (S.1) for $n=M$. Hence, on the basis of proposition 1, we can reformulate equation (S.13) as

$$S(\mathcal{A}) = \sum_{k=1}^M \left[\sum_{i=0}^{M-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=k}} S(\mathcal{I}).$$

From the following well-known property of binomial coefficients: $\sum_{j=0}^m (-1)^j \binom{m}{j} = 0$

$\forall m \in \{1, 2, \dots\}^2$, it turns out that all the coefficients in the squared brackets are zero unless $k=M$.

Then we can expand the right side of the last equation as follows:

$$S(\mathcal{A}) = 0 \cdot \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=1}} S(\mathcal{I}) + 0 \cdot \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=2}} S(\mathcal{I}) + \dots + 1 \cdot \sum_{\substack{\mathcal{I} \subset \mathcal{A} \\ |\mathcal{I}|=M}} S(\mathcal{I}).$$

As there is only one subset of \mathcal{A} with cardinality M , which is \mathcal{A} itself, the last expression is always the identity $S(\mathcal{A}) = S(\mathcal{A})$. Therefore proposition 2 is true.

Proposition 3. *The following expression*

$$S_{\mathcal{L}}^{(n)} = \sum_{k=1}^n (-1)^{k-1} \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} I_k(A_i, \mathcal{J}) \quad (\text{S.14})$$

is equal to

$$S_{\mathcal{L}}^{(n)} = \sum_{k=1}^n \sum_{i=1}^M \left[\sum_{j=0}^{n-k} (-1)^j \binom{|\mathcal{L}_i| - k}{j} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\} \quad (\text{S.15})$$

where $S_{\mathcal{L}}^{(n)}$ is our approximation to the R -dependant entropy $S_R^{(n)}$, being \mathcal{L} the non-redundant neighbor list of \mathcal{A} for the cutoff R (see the main text for further details). The \mathcal{J} index runs over all possible subsets of $(\mathcal{L}_i - \{A_i\})$ with $k-1$ elements.

Proof. We will follow a similar inductive reasoning to that used in the case of proposition 1. Note first that it can be shown that Proposition 3 is true for $n=1$ by taking into account that, in this particular case, $|\mathcal{J}|=0$, and hence $\mathcal{J} = \emptyset$. Rewriting (S.15) for $n+1$ we obtain:

$$S_{\mathcal{L}}^{(n+1)} = \sum_{k=1}^{n+1} \sum_{i=1}^M \left[\sum_{j=0}^{n-k+1} (-1)^j \binom{|\mathcal{L}_i| - k}{j} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\}. \quad (\text{S.16})$$

Next we split the outer sum by extracting the last term ($k = n+1$):

$$\begin{aligned} S_{\mathcal{L}}^{(n+1)} &= \sum_{k=1}^n \sum_{i=1}^M \left[\sum_{j=0}^{n-k+1} (-1)^j \binom{|\mathcal{L}_i| - k}{j} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\} \\ &\quad + \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\}. \end{aligned}$$

Similarly, we split the third most inner sum in the first term when $i = n - k + 1$:

$$\begin{aligned}
S_{\mathcal{L}}^{(n+1)} &= \sum_{k=1}^n \sum_{i=1}^M \left[\sum_{j=0}^{n-k} (-1)^j \binom{|\mathcal{L}_i| - k}{j} + (-1)^{n-k+1} \binom{|\mathcal{L}_i| - k}{n-k+1} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\} \\
&\quad + \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\}.
\end{aligned}$$

Under the hypothesis that our proposition is true for n , the last equation can be rewritten as

$$\begin{aligned}
S_{\mathcal{L}}^{(n+1)} &= S_{\mathcal{L}}^{(n)} + \sum_{k=1}^n \sum_{i=1}^M \left[(-1)^{n-k+1} \binom{|\mathcal{L}_i| - k}{n-k+1} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\} \\
&\quad + \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\},
\end{aligned}$$

where the two separate sums can be regrouped obtaining

$$S_{\mathcal{L}}^{(n+1)} = S_{\mathcal{L}}^{(n)} + \sum_{k=1}^{n+1} \sum_{i=1}^M \left[(-1)^{n-k+1} \binom{|\mathcal{L}_i| - k}{n-k+1} \right] \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \{S(A_i \cup \mathcal{J}) - S(\mathcal{J})\}. \quad (\text{S.17})$$

To demonstrate the validity of proposition 3 for $n+1$, we will derive an equivalent expression to Eq.(S.17), but transforming Eq.(S.14). Basing on the definition of $I_k(\mathcal{J})$ in Eq.(S.2), we can affirm that:

$$I_k(A_i, \mathcal{J}) = I_k(A_i, J_1, \dots, J_{k-1}) = \sum_{l=1}^k (-1)^{l+1} \sum_{\substack{\mathcal{I} \subset \{A_i, J_1, \dots, J_{k-1}\} \\ |\mathcal{I}|=l}} S(\mathcal{I}), \quad (\text{S.18})$$

which allows us to rewrite (S.14) for $n+1$ in terms of entropies instead of mutual information functions:

$$S_{\mathcal{L}}^{(n+1)} = \sum_{k=1}^{n+1} \sum_{i=1}^M \sum_{l=1}^k (-1)^k \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=k-1}} \sum_{\substack{\mathcal{I} \subset \{A_i, J_1, \dots, J_{k-1}\} \\ |\mathcal{I}|=l}} (-1)^l S(\mathcal{I}).$$

Next we split the outer sum in the last expression by extracting the corresponding term when $k=n+1$. The dummy index k vanishes after the sum splitting, but in the resulting expression we reintroduce k replacing the dummy index l for convenience:

$$S_{\mathcal{L}}^{(n+1)} = S_{\mathcal{L}}^{(n)} + \sum_i^M \sum_{k=1}^{n+1} (-1)^{n+1} \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=n}} \sum_{\substack{\mathcal{I} \subset \{A_i, J_1, \dots, J_n\} \\ |\mathcal{I}|=k}} (-1)^k S(\mathcal{I}) \quad (\text{S.19})$$

The last term in (S.19) is the exact value of the $(S_{\mathcal{L}}^{(n+1)} - S_{\mathcal{L}}^{(n)})$ difference. Simultaneously, the last term in (S.17) would also be equal to $(S_{\mathcal{L}}^{(n+1)} - S_{\mathcal{L}}^{(n)})$ under the induction hypothesis. It follows that the proof of proposition 3 implies that (S.17) and (S.19) are equivalent, what can be proven by transforming (S.19) into (S.17).

Let us observe now that a given subset \mathcal{I} can either include A_i or not. In any case, for every subset \mathcal{I} with cardinality k that includes A_i , there will be another subset with cardinality $k-1$ that does not include it. Then, we can rewrite the most inner sum in (S.19) running over the subsets of \mathcal{J} rather than $\{A_i, \mathcal{J}\}$ and including A_i explicitly as follows:

$$\begin{aligned} S_{\mathcal{L}}^{(n+1)} &= S_{\mathcal{L}}^{(n)} + \sum_i^M \sum_{k=1}^{n+1} (-1)^{n+1} \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=n}} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k-1}} \left[(-1)^k S(A_i \cup \mathcal{I}) + (-1)^{k-1} S(\mathcal{I}) \right] \\ &= S_{\mathcal{L}}^{(n)} + \sum_i^M \sum_{k=1}^{n+1} (-1)^{n+k+1} \sum_{\substack{\mathcal{J} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{J}|=n}} \sum_{\substack{\mathcal{I} \subset \mathcal{J} \\ |\mathcal{I}|=k-1}} [S(A_i \cup \mathcal{I}) - S(\mathcal{I})]. \end{aligned}$$

Finally, we merge the double sum over the subsets \mathcal{J} and \mathcal{I} into a single sum. To this end we need to count the number of times a given subset \mathcal{I} appears while summing over \mathcal{J} . The set \mathcal{J} has n elements and $(\mathcal{L}_i - \{A_i\})$ has $|\mathcal{L}_i| - 1$ elements, then once the $k-1$ elements of \mathcal{I} are selected, there remain $n-(k-1)$ unselected elements of \mathcal{J} . Thus, the number of times a given \mathcal{I} will appear is the

number of possibilities for selecting $n-(k-1)$ elements from the rest $(|\mathcal{L}_i|-1)-(k-1)$ elements,

which is $\binom{|\mathcal{L}_i|-k}{n-k+1}$. Therefore, the last expression can be transformed into:

$$S_{\mathcal{L}}^{(n+1)} = S_{\mathcal{L}}^{(n)} + \sum_{k=1}^{n+1} \sum_{i=1}^M \left[(-1)^{n-k+1} \binom{|\mathcal{L}_i|-k}{n-k+1} \right] \sum_{\substack{\mathcal{I} \subset (\mathcal{L}_i - \{A_i\}) \\ |\mathcal{I}|=k-1}} \{S(A_i \cup \mathcal{I}) - S(\mathcal{I})\} \quad (\text{S.20})$$

Substitution of the dummy ‘‘index’’ \mathcal{I} by \mathcal{J} makes (S.20) identical to (S.17). Thus proposition 3 is true.

Proposition4. *The Multibody Local Approximation (MLA; Eq. 8 in the main text) is exact for infinite cutoff.*

Proof. The MLA expression for a system $\mathcal{A} = \{A_1, \dots, A_M\}$ and cutoff R , is given by

$$S_{\mathcal{L}}(\mathcal{A}) = \sum_{i=1}^M \left[S(\mathcal{L}_i) - S(\mathcal{L}_i - \{A_i\}) \right], \quad (\text{S.21})$$

where $\mathcal{L}_i = \{A_j \in \mathcal{A} \mid j \geq i, d(A_i, A_j) < R\}$ (see the main text for the details).

For either an infinite cutoff or a cutoff greater than the system size, the lists \mathcal{L}_i can be constructed as follows:

$$\mathcal{L}_1 = \{A_1, \dots, A_M\}, \mathcal{L}_2 = \{A_2, \dots, A_M\}, \dots, \mathcal{L}_M = \{A_M\}.$$

Thus, $\mathcal{L}_{i+1} = \mathcal{L}_i - \{A_i\}$, being $\mathcal{L}_i = \emptyset$ if $i \notin \{1, 2, \dots, M\}$ and $S(\emptyset) = 0$. Now, Eq.(S.21) can be written as

$$S_{\mathcal{L}}(\mathcal{A}) = \sum_{i=1}^M [S(\mathcal{L}_i) - S(\mathcal{L}_{i+1})],$$

or, equivalently

$$S_{\mathcal{L}}(\mathcal{A}) = \sum_{i=1}^M S(\mathcal{L}_i) - \sum_{i=2}^M S(\mathcal{L}_i) = S(\mathcal{L}_1) = S(\mathcal{A}).$$

2. Reordering the elements of \mathcal{A} for minimizing the numerical effects of the permutation inconsistency in $S_{\mathcal{L}}^{(n)}$

As mentioned in the main text, $S_{\mathcal{L}}^{(n)}$ (Eq. 7) is not strictly invariant under any $A_i \leftrightarrow A_j$ permutation because $S_{\mathcal{L}}^{(n)}$ contains additional terms with respect to those included in $S_R^{(n)}$. Hence, we have

$$S_{\mathcal{L}}^{(n)} = S_R^{(n)} + \delta_{\mathcal{L}}^{(n)},$$

where

$$\delta_{\mathcal{L}}^{(n)} = \sum_{k=1}^n (-1)^{k-1} \sum_{i=1}^M \sum_{\substack{\mathcal{J} \subset \{\mathcal{L}_i - \{A_i\}\} \\ |\mathcal{J}|=k-1 \\ \mathcal{J} \in \{\mathcal{C}(2R) - \mathcal{C}(R)\}}} I_k(A_i, \mathcal{J})$$

collects the contributions to the entropy of those additional terms that belong to the $\mathcal{C}(2R) - \mathcal{C}(R)$ set. Unlike $S_R^{(n)}$, both $S_{\mathcal{L}}^{(n)}$ and $\delta_{\mathcal{L}}^{(n)}$ are not invariant under any $A_i \leftrightarrow A_j$ permutation. Since the number of these terms also depends on the ordering of the elements of $\mathcal{A} = \{A_1, \dots, A_M\}$, we can arrange a proper reordering of \mathcal{A} in order to minimize $\delta_{\mathcal{L}}^{(n)}$ and, consequently, the permutation inconsistency.

A simple algorithm that seeks to minimize the number of extra elements included in $\mathcal{D}_{\mathcal{L}}^{(n)}$ consists of the following steps:

1. For each element A_i , we count the number of subsets $\mathcal{I} \subset (\mathcal{L}_i - \{A_i\})$ that belong to $\mathcal{C}(2R) - \mathcal{C}(R)$ and have two elements (*i.e.*, $|\mathcal{I}| = 2$).
2. We select the A_i element that results in the minimum number of subsets as constructed in the step 1). If several elements fulfill this condition, we simply select the A_i element showing the minimum i -index as we have no reason to prefer any particular element.
3. The selected element in 2), for example A_k , is taken as the first element B_1 in a reordered set $\mathcal{B} = \{B_1, \dots\}$. Subsequently A_k is removed from the original set \mathcal{A} .
4. The same steps 1→2→3 are followed to select B_2, B_3, \dots and so on. Finally, the reordered set \mathcal{B} replaces \mathcal{A} during the entropy calculations.

Note, however, that this algorithm does not aim at finding a global minimum number of additional terms in $\mathcal{D}_{\mathcal{L}}^{(n)}$. In fact this simple strategy seeks to reduce the permutation inconsistency of $\mathcal{S}_{\mathcal{L}}^{(n)}$ at a low computational cost $O(M^2)$. Other more efficient strategies for performing a convenient reordering of the \mathcal{A} elements could be devised although a full minimization of $\mathcal{D}_{\mathcal{L}}^{(n)}$ would require to explore all the $M!$ permutations of the A_i elements, which in general would be prohibitively expensive.

3. Conformational Entropy: Testing the MLA Performance

Table S1. Conformational entropy values ($cal\ K^{-1}\ mol^{-1}$) and computer CPU time^(a) (s) in parentheses for $S_R^{(n)}$ and $S_{\mathcal{L}}$ (MLA) test calculations on 2000 MD snapshots extracted from the **f-THP5** simulation. The order of the expansion n goes from 1 to the maximum allowed (*Max*) order at the given cutoff.

n	$S_R^{(n)}(R = 8\text{\AA})$		$S_R^{(n)}(R = 9\text{\AA})$		$S_R^{(n)}(R = 10\text{\AA})$	
1	129.58	(0.1)	129.58	(0.07)	129.58	(0.14)
2	57.74	(0.24)	52.87	(0.24)	44.65	(0.29)
3	77.64	(0.63)	71.72	(1.01)	58.43	(1.84)
4	81.07	(2.07)	77.32	(4.79)	73.47	(12.32)
5	79.48	(5.66)	74.43	(18.70)	65.65	(60.69)
6	77.94	(12.39)	72.87	(56.50)	63.52	(256.27)
7	79.27	(22.05)	74.98	(142.28)	69.30	(919.02)
8	78.85	(33.57)	73.74	(304.96)	62.96	(2872.66)
9	78.91	(45.05)	74.21	(576.06)	68.80	(8069.77)
10	78.90	(55.08)	74.10	(976.49)	64.79	(20581.72)
11	78.91	(62.19)	74.12	(1480.89)	66.51	(47989.23)
12	78.90	(66.75)	74.11	(2059.34)	66.16	(113634.60)
13	78.90	(68.92)	74.11	(2625.95)	66.12	(261403.28)
14	78.90	(69.84)	74.11	(3120.65)	66.17	(445027.59)
15	78.90	(70.20)	74.11	(3190.85)	66.15	(717127.62)
16	78.90	(70.14)	74.11	(3309.65)	66.16	(1074472.62)
Max	78.90	(70.14)	74.11	(3572.72)	66.16	(1909558.38)
MLA	78.83	(0.44)	74.07	1.23	66.40	1.24

^(a)Using a Xeon 5360 core.

4. Conformational Entropy of fTHP-5 using the Classical Mutual Information Expansion

Although the classical MIE is exact for *maximum* order as pointed out in the main text and in Proposition 2, it is clear that practical MIE calculations at low orders suffer to some extent from truncation errors. Unfortunately, truncation errors at third or larger orders behave erratically, what precludes us from selecting *a priori* a value of n for achieving a given accuracy. On the other hand, the large majority of the MIE applications have been limited to second or third order approximations in order to keep the computational cost within tractable limits as well as to obtain reasonably converged mutual information functions with low bias. Even in the case that we might afford the computational cost of high order MIE calculations, they would become unreliable given that, for a finite amount of sampling, MIE would converge towards the exact value of the *sample* entropy, that is, to a *biased* entropy value. All these problems and limitations become particularly acute for large molecular systems like the fTHP-5 peptide, which has 172 rotatable bonds, as shown by the data collected in Table S2 and Figure S1.

The first order approximation to the conformational entropy $S^{(1)}$ is the sum of marginal entropies and therefore constitutes an upper bound to the actual entropy. Consequently, since the entropy of a discrete variable is always positive, the “exact” entropy for the given sample must lie within the $[0, S^{(1)}]$ interval and, in the worst scenario, the absolute error of $S^{(1)}$ would amount to $S^{(1)}$ itself. For the fTHP-5 system, we obtained quite well converged values of $S^{(1)}$ after having accumulated molecular configurations every ps for at least 600 ns of MD simulation. In contrast Table S2 and Figure S1 reveal that the second order value $S^{(2)}$ is negative all along the simulation time while the third order approximation $S^{(3)}$ gives positive entropy values that are much larger than $S^{(1)}$.

Certainly, the corresponding $S^{(2)}$ and $S^{(3)}$ plots versus simulation time exhibit a poorer convergence than the first order entropy (see Figure S1). It may also be interesting to note that, in this particular case, the $S^{(3)}$ values do not give better entropy estimations than the $S^{(1)}$ ones because $S^{(3)}$ is more than twice $S^{(1)}$ and, in the best case, the absolute error of $S^{(3)}$ amounts to $S^{(3)} - S^{(1)}$ (larger than $S^{(1)}$). In other words, the largest error in the $S^{(1)}$ values is below the smallest error in the $S^{(3)}$ data regardless the exact value of the fTHP-5 conformational entropy.

Clearly, the observed instability of the classical MIE for large systems restricts its applicability to smaller systems for which the *combinatory* is moderate. Otherwise, insufficient sampling and truncation errors with respect to the order expansion can determine that even low order calculations become unreliable. It is important to emphasize that, in general, the influence of truncation errors at various orders cannot be neglected by increasing the sampling effort. To show more clearly this point, let us suppose that, starting from a finite MD simulation of a *real* biomolecule, we replicate it in order to build a hypothetical system in which the internal degrees of freedom have the same probability density functions as in the simulated (real) system, but have a periodic time evolution with a period equal to the simulation time of the original system. Since the hypothetical imaginary system is periodic in time, its MIE entropy values $S^{(1)}, S^{(2)}, S^{(3)} \dots$ could be computed exactly as the underlying probability density functions would also be exact using data from one period. Nevertheless, although the MIE entropy calculations for the hypothetical system would not have any statistical bias, they would still exhibit convergence problems due to truncation errors similar to those of the real system, requiring a generally unaffordable cost for increasing the expansion order.

Figure S1. Plot of the convergence of the conformational entropy estimation for fTHP5 using orders 1, 2 and 3.

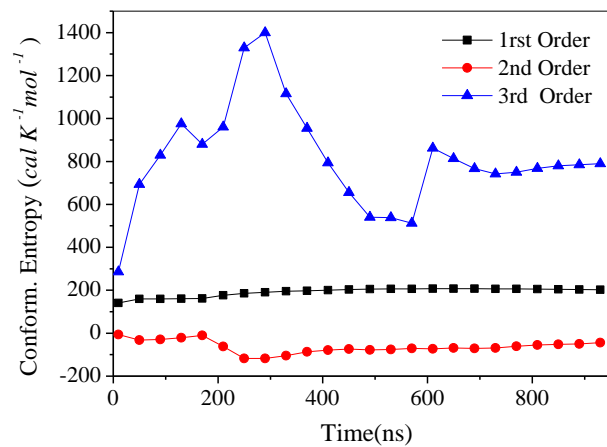


Table S2: Conformational entropy ($\text{cal K}^{-1} \text{mol}^{-1}$) estimations for fTHP-5 at various simulation times (ns) using the classical MIE up to third order.

Time	S^{conform}		
	$S^{(1)}$	$S^{(2)}$	$S^{(3)}$
10	141.24	-6.68	285.71
50	159.54	-32.02	693.09
90	159.68	-29.07	829.59
130	160.31	-20.90	975.64
170	162.11	-9.96	880.49
210	176.93	-62.32	960.18
250	185.81	-117.18	1329.36
290	190.33	-117.62	1399.22
330	195.07	-104.49	1115.27
370	196.95	-86.96	954.57
410	200.42	-78.32	793.87
450	203.57	-74.05	655.77
490	205.26	-78.23	540.20
530	206.31	-75.73	538.12
570	206.83	-71.01	512.67
610	207.84	-72.60	862.62
650	207.71	-69.17	813.27
690	207.22	-70.97	766.25
730	206.55	-68.78	741.93
770	205.89	-61.30	749.62
810	205.10	-55.29	768.22
850	204.08	-52.39	779.33
890	203.06	-49.90	784.46
930	202.27	-44.28	789.37

References

- (1) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 2011, 7, 2638-2653.
- (2) Comtet, L. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*; Reidel Publishing Company: Holland, 1974.

**2.1.1.6 CENCALC: A New Program for Conformational Entropy Calculation
of Macromolecules from Molecular Dynamics Simulation**

Ernesto Suárez, Natalia Díaz, Jefferson Méndez, and Dimas Suárez
(Enviado para su publicación a *J. Chem. Inform. Model.*)

CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations

Journal:	<i>Journal of Chemical Information and Modeling</i>
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Suarez, Ernesto; Universidad de Oviedo, Quimica Fisica y Analitica Diaz, Natalia; Universidad de Oviedo, Quimica Fisica y Analitica Mendez, Jefferson; Universidad de Oviedo, Quimica Fisica y Analitica Suarez, Dimas; Universidad de Oviedo, Quimica Fisica y Analitica

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations

Ernesto Suárez, Natalia Díaz, Jefferson Méndez and Dimas Suárez**

Departamento de Química Física y Analítica.

Universidad de Oviedo.

C/ Julián Clavería, 8. 33006, Oviedo. Spain.

dimas@uniovi.es

ernesto@fluor.quimica.uniovi.es

TELEPHONE: +34-985103689

FAX: +34-985103125

ABSTRACT

Herein, we present the CENCALC software that has been designed to estimate the conformational entropy of single molecules from extended Molecular Dynamics (MD) simulations. CENCALC uses both trajectory coordinates and topology information in order to characterize the conformational states of the molecule of interest by discretizing the time evolution of internal rotations. Subsequently, entropies can be obtained using various techniques. On one hand, CENCALC can approach to the full conformational entropy by means of the mutual information expansion (MIE), which is built upon the converged probability density functions of the individual torsion angles, pairs of torsions, triads and so on. In addition, CENCALC implements various reformulations of MIE including the so-called Multibody Local Approximation (MLA), which captures implicitly all-order correlation effects within a given cutoff. A correlation-corrected entropy estimator is also implemented that allows users to select the best cutoff for the MLA calculations in order to retrieve the maximum amount of genuine correlation from a given MD trajectory. The CENCALC software, which comprises two FORTRAN90 codes complemented with an auxiliary Python script that provides an interface with the Amber package of programs, can be executed in parallel computers using the OpenMP technology for shared-memory architectures. The software, which is distributed under the GNU public license, together with numerical examples and a user's manual are available in the Supporting Information and at the URL <http://sourceforge.net/projects/cencalc/>.

INTRODUCTION

The estimation of absolute entropies from molecular simulations has been a topic of great interest in Computational Chemistry during the last decades.¹⁻² This problem is particularly relevant for the theoretical analysis of large and flexible biomolecules that populate multiple energy wells on their potential energy surfaces. Although different approaches have been proposed for analyzing the role of entropy in the stability and behaviour of these systems,³⁻⁷ the estimation of their absolute entropies is still a challenging problem. As a matter of fact, many questions concerning the influence of correlation among internal motions remain open given that, for relatively large molecular systems, the majority of entropy estimations have made use of linear approximations or low-order truncated mutual information expansions.^{3,8-9}

Many review and research articles have been published in the literature in which the various methodologies for configurational and/or conformational entropy calculations of single molecules are described in detail.^{1,3-8,10-16} As the main goal of the present paper is to present the capabilities of the CENCALC (*Conformational ENTropy CALCulations*) software, readers who are interested rather in formulation of the entropy methods should consult with the original references that are cited throughout the text. Herein, we briefly comment about a few methods that are related with the techniques implemented in CENCALC. Most of them aim at the evaluation of the following multidimensional integral, which defines the classical configurational entropy:

$$S_{config} = -R \int P(\mathbf{q}) \ln P(\mathbf{q}) d\mathbf{q}$$

where R is the ideal gas constant, \mathbf{q} represents one of the configurations of the system, and $P(\mathbf{q})$ stands for the configurational probability density function (PDF) in an N -dimensional

1
2
3 configurational space. The so-called *parametric* methods assume *a priori* a functional form for
4 $P(\mathbf{q})$, and thereby the computational task is formally reduced to a point estimation problem. This
5
6
7
8 is the case of the broadly used Quasi Harmonic (QH) methods, which approximate $P(\mathbf{q})$ as a
9
10 multivariate normal distribution.¹²⁻¹³ However, the QH methods exhibit two well known
11
12 disadvantages: (a) supralinear correlations among the system variables are neglected, and (b) the
13
14 actual multimodal PDF is approximated by a monomodal one. The consequence is an important
15
16 non-systematic overestimation of S_{config} ,¹⁷ which is also due to the fact that, for a given
17
18 covariance matrix, the function that maximizes entropy is precisely the normal distribution
19
20 function.¹⁸ Interestingly, other authors have proposed methods that refine the QH entropies *a*
21
22 *posteriori* by computing correction terms accounting for anharmonicity and supralinear
23
24 correlation effects.⁴⁻⁵
25
26
27
28
29

30 A second group of direct entropy methods are *non-parametric* in the sense that they seek
31
32 to estimate $P(\mathbf{q})$ without resorting to any analytical approximation. Of particular interest for the
33
34 methods that are implemented in CENCALC is the former work done by Gilson and co-
35
36 workers.^{3,8} These authors compute the total entropy as an expansion of mutual information terms,
37
38 which take into account the correlated motions among the internal degrees of freedom of the
39
40 molecule. This expansion, also called Mutual Information Expansion (MIE),⁹ is equivalent to
41
42 estimating the global PDF using the generalized Kirkwood superposition approximation.³
43
44 Unfortunately, the Gilson's method has been applied to relatively small molecular systems or, in
45
46 general, to low-dimensional systems.^{8,19}
47
48
49
50
51

52 Concerning the methods implemented in CENCALC, they could be classified as non-
53
54 parametric because they do not make any assumption about the PDF of single molecules.
55
56 However, it is important to remark that CENCALC estimates only the *conformational* entropy,
57
58
59
60

1
2
3 which arises from the partitioning of the total entropy (S_{tot}) of a single molecule (excluding
4 translation and rotation) into a vibrational (\bar{S}_{vib}) and a pure conformational contribution ($S_{conform}$):
5
6
7

$$S_{tot} = \bar{S}_{vib} + S_{conform}$$

8
9
10
11
12 This decomposition, which was first proposed by Karplus and co-workers,¹⁰⁻¹¹ can be shown to
13 be exact for a molecular system provided that its potential energy surface can be described as a
14 collection of separate energy wells.²⁰⁻²¹ As shown in previous work, this entropy decomposition
15 provides quite accurate results for small molecules in the gas phase (ideal conditions) at room
16 temperature.²¹⁻²² Although the distinction between vibrational and conformational motions
17 would become blurred as temperature increases (due to the occupation of high energy vibrational
18 levels that can be simultaneously assigned to different conformers), the decomposition of S_{tot}
19 would result in an upper bound to the actual entropy value.
20
21
22
23
24
25
26
27
28
29
30
31

32
33 The practical implementation of the entropy decomposition within the context of
34 Molecular Dynamics (MD) simulations of biomolecules has already been demonstrated in
35 previous works.^{21,23} On one hand, it has been shown that the average value of S_{vib} can be readily
36 computed over a representative set of structures extracted from the corresponding MD trajectory
37 by means of energy minimization calculations followed by classical normal mode analyses using
38 either quantum mechanical (QM) or molecular mechanical (MM) methods. Basically, the same
39 procedure is employed by approximate end-point free energy methods like the Molecular
40 Mechanics Poisson-Boltzmann (MM-PB) protocol.²⁴ On the other hand, the $S_{conform}$ term has been
41 obtained separately by determining first the conformational states along the MD trajectory
42 through the discretization of the time evolution of internal rotations about rotatable bonds.
43 Subsequently, $S_{conform}$ is estimated by means of either the MIE method up to a given order or any
44 of the MIE reformulations that capture correlation effects within a given cutoff (see below). Note
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 also that the addition of $\bar{S}_{vib} + S_{conform}$ to the rotational-translational entropies provided by the
4
5 Rigid Rotor & Harmonic Oscillator (RRHO) model together with the standard formulae of
6
7 Statistical Thermodynamics, result in absolute entropies that in some cases admit a direct
8
9 comparison with experimental data (e.g., in the gas phase).²¹⁻²²
10
11

12
13
14 In this work, we will be basically concerned with the calculation of $S_{conform}$ as defined by
15
16 the $S_{tot} = \bar{S}_{vib} + S_{conform}$ entropy partitioning using the various techniques implemented in
17
18 CENCALC. Thus, all the options and capabilities of CENCALC will be shown throughout a
19
20 series of conformational entropy calculations for a decapeptide molecule that are performed over
21
22 10^6 configurations extracted from a 1.0 μ s MD simulation. The corresponding peptide sequence
23
24 has been selected from a peptide library for probing the cleavage site motifs of matrix
25
26 metalloproteinases (MMPs).²⁵ Particular attention will be paid to the calculations performed in
27
28 the framework of the so-called Multibody Local Approximation (MLA)²⁶ and the determination
29
30 of the best cutoff criterion for a given amount of MD sampling. Besides analyzing the
31
32 convergence properties of the entropy plots as a function of the simulation time and/or the
33
34 expansion order in the MIE calculations, we will also discuss the potential utility of the $S_{conform}$
35
36 values for analyzing the dynamical properties of a single molecule and its entropy changes
37
38 occurring upon molecular association processes. To this end, we will also examine the binding of
39
40 the selected decapeptide with the MMP-7 enzyme²⁷ by carrying out an extended MD simulation
41
42 followed by conformational entropy calculations over the coordinates of the peptide molecule.
43
44 Overall, we feel that the approximate entropy methods implemented in CENCALC could be
45
46 useful for studying polypeptide-folding or ligand binding processes.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

MATERIALS AND METHODS

MD settings

Simulation of the isolated GNR peptide in aqueous solution

The examined model peptide, which will be named hereafter as **GNR**, is an active substrate for the recombinant form of stromelysin (MMP-7).²⁸ It contains the following amino acid sequence: *Ace-Gly-Ile-Pro-Phe-Glu~Gln-Arg-Leu-Val*. The starting structure for the MD simulation was generated with the help of the LMOD algorithm included in the Amber10 package²⁹⁻³⁰ using the *parm03* force field³¹ coupled with the Hawkins-Cramer-Truhlar (HCT) pairwise Generalized-Born (GB) solvent model.³² A total of 1000 LMOD iterations were computed by exploring one-frequency vibrational modes where all the peptide residues were allowed to move, computing the eigenvectors every 25 LMOD iterations.

The structure with the lowest LMOD energy was selected as the initial point for the MD simulation. The solute molecule was surrounded by a truncated octahedral periodic box of TIP3P water molecules that extended 12 Å beyond the peptide atoms. In addition, a counter ion (Na⁺) was placed at one of the edges of the box to neutralize the system, resulting in a total of 158 peptide atoms being solvated by 1830 water molecules under periodic boundary conditions.

Energy minimizations and MD simulations were carried out using the SANDER and PMEMD programs included in the Amber10 suite of programs.^{30,33} The SHAKE algorithm was used to constraint all the R-H bonds. A nonbonded cutoff of 10.0 Å was employed, whereas the Particle-Mesh-Ewald (PME) method was used to include the contributions of long-range interactions. The pressure (1 atm) and the temperature (300 K) of the system were controlled during the MD simulations by Berendsen's method during a total simulation time of 1.0 μs using

1
2
3 a time step of 2 fs. The coordinates were saved for analysis and conformational entropy
4
5 calculations every 1 ps.
6
7
8
9

10 11 *Setup of the MMP-7/GNR simulation*

12
13
14
15 Initial coordinates for the MMP-7 enzyme were taken from the 1MMQ crystal structure (1.90 Å)
16
17 that corresponds to an inactive form of the enzyme due to the presence of an hydroxamate
18
19 inhibitor bound to the catalytic zinc ion in the active site.³⁴ In this crystal structure, the *N*-
20
21 terminal tail of the catalytic domain is placed as in the so-called "superactivated form",³⁵ a
22
23 positioning that was maintained in our model. In addition, we also included the two zinc and two
24
25 calcium ions bound to the catalytic domain that are observed in the 1MMQ structure, while the
26
27 ionizable residues were set to their normal ionization states at pH 7.
28
29
30
31
32
33
34
35

36 To obtain an initial structure for the complex formed between MMP-7 and **GNR**, we
37
38 manually transformed the inhibitor atoms that interact with the enzyme in the 1MMQ structure to
39
40 their counterparts in the **GNR** peptide. The remaining of the **GNR** backbone was built by
41
42 superimposing the partially-transformed MMP-7 structure onto a representative structure of the
43
44 complex formed between the MMP-2 catalytic domain and a similar peptide molecule that was
45
46 taken from a previous simulation.³⁶ The **GNR** side chains were finally added by the LeAP
47
48 program included in the Amber10 package.³⁰ The resulting MMP-7/GNR structure was then
49
50 solvated by all the crystallographic water molecules, except those that collapsed with the newly
51
52 created substrate, and by a rectangular solvent box that extended 15 Å from the protein atoms. In
53
54
55
56
57
58
59
60

1
2
3 addition, five Cl^- counterions were added by LeAP in order to neutralize the system (this resulted
4
5
6 in a system containing a total of 38832 atoms).
7

8
9 The *parm03* version of the all-atom AMBER force field was used to model the solvated
10
11 MMP-7/GNR complex.³¹ For the metal ions, we used an MM representation previously
12
13 developed and tested for the MMP-2 enzyme.³⁷ The peptide substrate and all the solvent
14
15 molecules were initially minimized to remove bad contacts using the SANDER program in
16
17 Amber10. Subsequently, the solvent molecules and counterions were relaxed by 100 ps of MD.
18
19 The full system was then minimized and gradually heated to 300 K during 60 ps. Finally, we
20
21 computed a 400 ns trajectory using the same settings already described for the simulation of the
22
23 unbound **GNR** substrate in aqueous solution. In this case, we employed the GPU accelerated
24
25 version of the PMEMD code included in Amber 11³⁸ using the recommended SPDP precision
26
27 model, which provides an optimum tradeoff between accuracy and performance.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

CENCALC: A brief description

In principle, CENCALC can process either a single MD trajectory or the accumulated information from a series of independent MD runs. Similarly, conformational entropies could also be obtained from data generated by Metropolis Monte Carlo simulations. Nevertheless, and for the sake of simplicity, we will assume in the following description that CENCALC processes a single MD trajectory.

The CENCALC software consists mainly of two independent codes written in the FORTRAN90 language, *cencalc_prep.f90* and *cencalc_omp.f90*. The first program carries out various preparatory tasks prior to the main entropy calculations that are performed by *cencalc_omp.f90*. As the bulk of the effort to obtain the conformational entropy is expended by *cencalc_omp.f90*, this program was designed to take advantage of shared-memory parallel computers through the OpenMP Application Program Interface.

In addition, CENCALC also includes a *Python* program (*get_tor.py*) that reads the topology information from Amber *parm* files and identifies all the torsions about rotatable bonds that are required to characterize the conformational state of the molecule of interest. Command line options of *get_tor.py* allow users to select different subsets of torsion angles for entropy analyses (e.g., main chain torsions of polypeptide molecules). On output *get_tor.py* produces an input file for the program *ptraj*,³³ which is a general purpose utility for analyzing and processing MD trajectory files included in the freely distributed *AmberTools* package. Thus, execution of *ptraj* results in a series of data files containing the evolution of the selected torsion angles all along the MD trajectory and an averaged distance matrix among the solute atoms. We note,

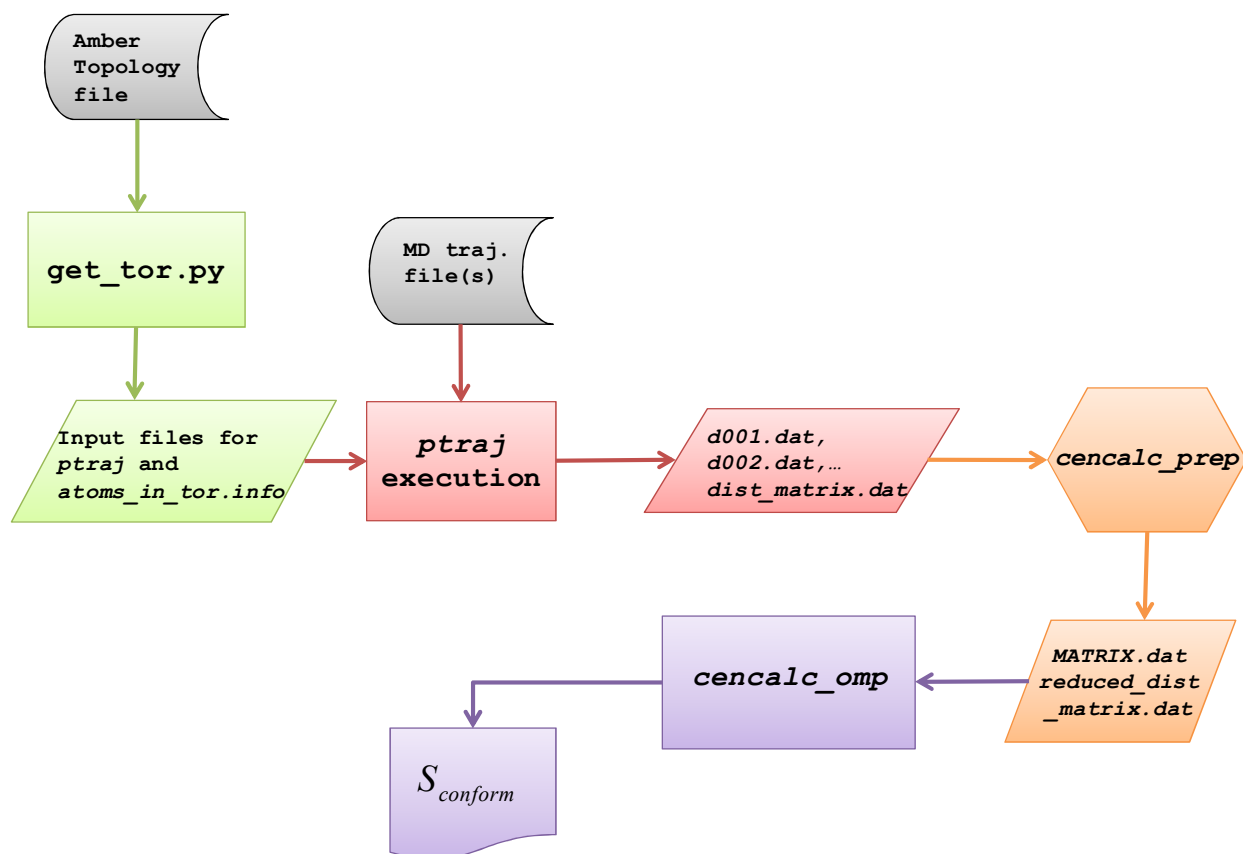
1
2
3 however, that users of other simulation packages could easily develop alternative protocols to
4 generate all the data to be processed by *cencalc_prep.f90* and *cencalc_omp.f90*.
5
6
7

8
9 Figure 1 shows a typical flowchart for using sequentially the CENCALC components.
10 Starting with one MD trajectory and its corresponding topology file, *get_tor.py* generates the
11 input file for *ptraj* and the *atoms_in_tor.info* file, the latter one being required in entropy
12 estimations using a distance-based cutoff. Execution of *ptraj* or other alternative computational
13 tools results in a text file per torsion angle containing the corresponding time series (default
14 names: *d0001.dat*, *d0002.dat*, ...) as well as the interatomic mean distance matrix if required.
15
16 The torsion files are processed by *cencalc_prep.f90*, which removes all the frozen torsions and,
17 most importantly, transforms the initial time series containing N data points $\{\theta_1, \dots, \theta_N\}$ per
18 torsion angle θ into a set of N integer numbers $\{x_1, \dots, x_N\}$ labeling the conformational states
19 populated by the torsion angle (see below). The resulting integer arrays are then collected into a
20 rectangular matrix (default name *MATRIX.dat*) with dimensions $N \times M$, where M is the number of
21 rotatable bonds and N the number of snapshots. In addition, the preparatory program,
22 *cencalc_prep.f90*, reads the *atoms_in_tor.info* file and reduces the interatomic mean distance
23 matrix to give a new matrix $M \times M$, which collects the mean value of distances among the center
24 of mass of the two central atoms that are involved in each one of the M torsion angles.
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

46 Before performing the entropy calculations, the *cencalc_omp.f90* program first reads the
47 input data (*MATRIX.dat* and *reduced_dist_matrix.dat*) and then calculates the probability mass
48 function of the torsion angles by means of the maximum likelihood method fed with the
49 corresponding outcomes $\{x_1, \dots, x_N\}$ along the MD simulation. Next the program typically
50 computes data points for obtaining an entropy plot with a time step specified in terms of a
51 number of MD snapshots and using one of the implemented entropy methods (see below). At this
52
53
54
55
56
57
58
59
60

1
2
3 stage, *cencalc_omp.f90* builds on-the-fly all the subsets of torsion angles that fulfill the
4 corresponding distance-based cutoff criteria (if required) and/or that contain up to n elements if
5 the MIE technique is used at order n . During the runtime of *cencalc_omp.f90*, the program prints
6 out periodically information about the progress of the calculations. On output, a text file is
7 generated that contains a Table summarizing all the calculations and that can be readily imported
8 by spreadsheet software for data analysis. Further information and technical details about the
9 usage of CENCALC can be found in the users' manual that is included in the Supporting
10 Information.

21
22
23
24
25
26 **Figure 1.** Flowchart for obtaining conformational entropies using the CENCALC software (See
27 text for the details).



Discretization of the Torsional Degrees of Freedom

As above mentioned, the conformational states of a single molecule can be characterized by means of the discretization of the time evolution of internal rotations. Although the discretization algorithm implemented in CENCALC has been described in detail in a previous work,²¹ herein we provide again a brief description of this basic transformation in order to show a global view of the software. For each torsion angle, the continuous PDF is first estimated from a set of N outcomes extracted along the MD simulation $\{\theta_1, \dots, \theta_N \mid \theta_i \in [0, 2\pi)\}$ using the von Mises kernel density estimation given by

$$\hat{\rho}(\theta; \nu) = \frac{1}{2\pi N I_0(\nu)} \sum_{i=1}^N \exp\{\nu \cos(\theta - \theta_i)\}$$

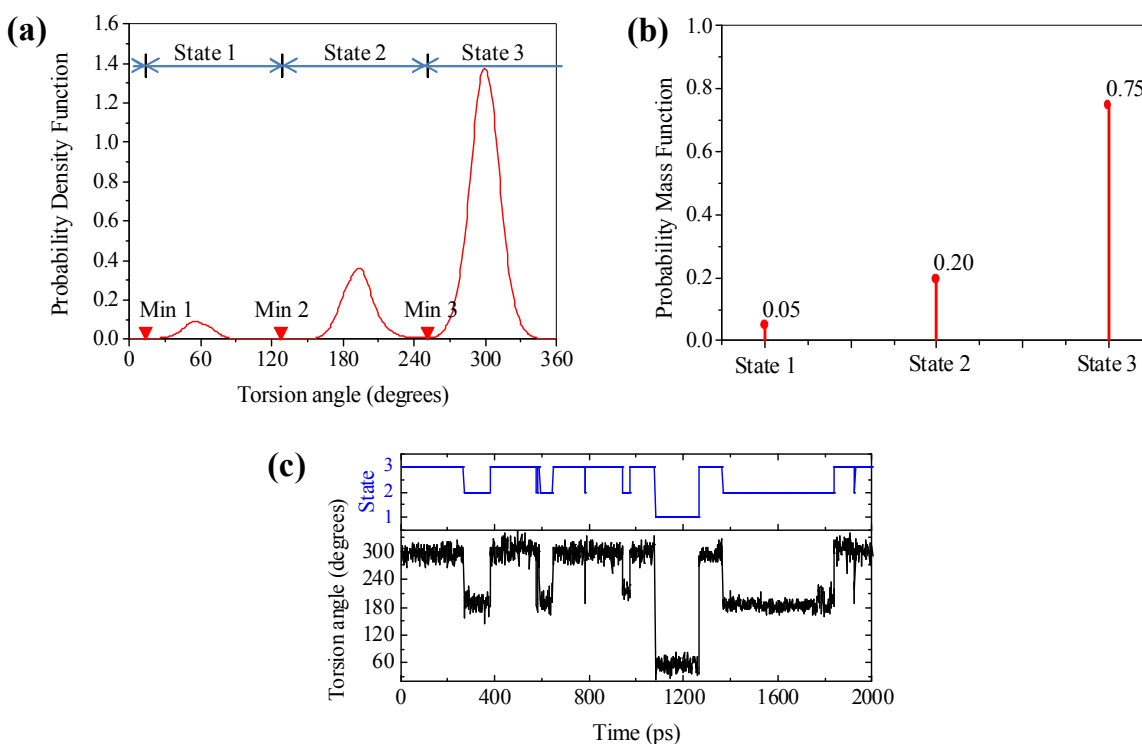
where $I_r(\nu)$ is the modified Bessel function of order r , and ν is the so-called concentration parameter. The value of ν is obtained by applying the recently derived “von Mises-scale plug-in rule”,³⁹ the smoothing parameter being computed by the following expression that depends on the number N of data points:

$$\nu = \left[3N\hat{\kappa}^2 I_2(2\hat{\kappa}) \left\{ 4\pi^{1/2} I_0(\hat{\kappa})^2 \right\}^{-1} \right]^{2/5}$$

where $\hat{\kappa}$ is an estimation of the concentration parameter of the global data. CENCALC takes by default $\hat{\kappa}=1$ and searches for the critical points of the resulting distribution using both its first and second derivatives ($\hat{\kappa}=1$ ensures that $\hat{\rho}(\theta; \nu)$ is a slightly oversmooth function, what is convenient for locating its critical points). After having found the subintervals delimiting the

1
2
3 conformational states accessible to each torsion angle as graphically shown in Figure 2,
4
5 CENCALC transforms the initial time series containing N data points, $\{\theta_1, \dots, \theta_N\}$, into a set of N
6
7 integer numbers $\{x_1, \dots, x_N\}$ labelling the conformational states populated by the torsion angle. In
8
9 this way, the continuous variable θ becomes a discrete random variable X with probability mass
10
11 function $P(X)$ (see Figure 2). Note that the actual labels used to specify the conformational states
12
13 are irrelevant given that the entropy only depends on the probability (*i.e.*, the observed frequency)
14
15 of those labels.
16
17
18
19
20
21
22
23
24

25 **Figure 2.** (a) Probability Density Function for the torsion angle θ defined by the C γ 2-C β -C γ 1-C δ 1
26
27 atoms of the Ile₃ residue of **GNR** as obtained from a von-Mises kernel estimator. (b) Probability
28
29 mass function of the three conformational states. (c) Time evolution of the torsion angle θ and its
30
31 associated discrete variable (see text for details).
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

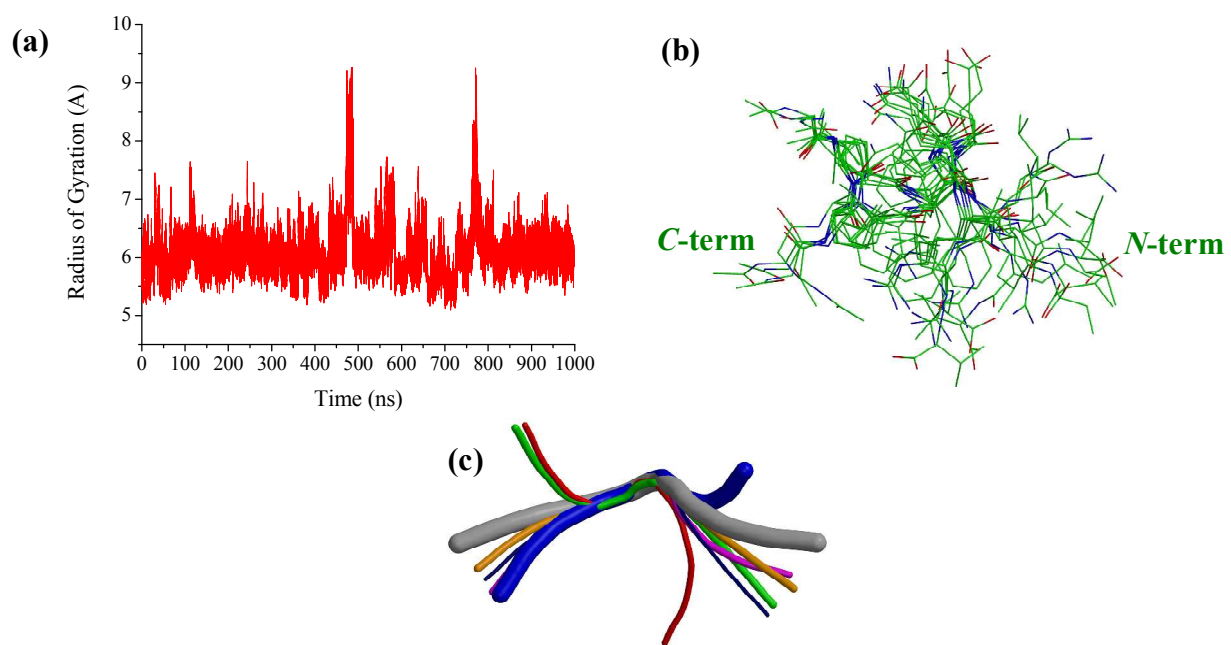


MD simulation of the isolated form of GNR

The 1.0 μs MD trajectory of the **GNR** peptide in explicit solvent populates mainly a conformational region characterized by a radius of gyration (r_{gyr}) of $\sim 6.0 \pm 0.4 \text{ \AA}$. The r_{gyr} plot also reveals two conformational changes at 480 and 770 ns of simulation time, respectively, that take place in less than 15 ns through extended structures with $r_{gyr} \sim 8\text{-}9 \text{ \AA}$. On the other hand, secondary structure analyses assign a helical conformation to the central residues in $\sim 70\%$ of the analyzed snapshots while, as expected, the terminal residues are mainly random coils. This quasi-helical conformation is stabilized by intra-molecular H-bond interactions involving backbone groups of the central residues (*e.g.* Pro₄-C=O \cdots HN-Gln₇, Pro₄-C=O \cdots HN-Arg₈, Phe₅-C=O \cdots HN-Arg₈, Ile₃-C=O \cdots HN-Glu₆, etc.), which all have a 40-50% abundance along the MD simulation. Clustering analyses confirm that the **GNR** peptide exhibits a remarkable dynamical flexibility through either

1
2
3 its backbone or side chain motions as shown in Figure 3, in which the 8 most populated
4
5 representatives are superimposed onto each other. Hence, all these analyses suggest that
6
7 conformational entropy should result in a significant free energy contribution stabilizing the
8
9 unbound form of the **GNR** peptide in aqueous solution.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8 **Figure 3** (a) Radius of gyration along the MD simulation for the isolated **GNR** peptide. (b) and
9
10 (c) superposition of the representative structures accounting for 90% of backbone variability
11
12 according to clustering analyses, both in wireframe and ribbon model representations. Thickness
13
14 of the ribbon models corresponds to the number of snapshots represented by each model.
15
16
17
18
19



1
2
3
4 *Methods implemented in CENCALC for conformational entropy calculations:*

5
6
7 *Application to the isolated GNR peptide*

8
9
10 After the discretization of the torsion angle evolution, the conformational state of an individual
11 torsion becomes associated with a one-dimensional random variable X . Analogously, the
12 conformational state of a set of M torsion angles $\{A_1, \dots, A_M\}$ can be described by an M -
13 dimensional random variable or vector $X = (X_1, \dots, X_M)$, whose components specify the discrete
14 conformational state of the different torsions. In principle, one could estimate the conformational
15 entropy directly from the observed relative frequencies of the outcomes x of X by means of the
16 Shannon expression:
17
18
19
20
21
22
23
24
25
26

$$S_{conform}(A) = -R \sum_x \hat{p}_x \ln \hat{p}_x \quad (1)$$

27
28
29 where \hat{p}_x is an estimation, using the observed relative frequencies, of the probability to obtain
30 the outcome x . Unfortunately, the number of potentially accessible conformers for medium-sized
31 and large molecules is huge ($\sim 3^M$) and, therefore, the direct application of the Shannon
32 expression would result in large and negatively-biased entropies due to sampling limitations. In
33 what follows, we illustrate different strategies implemented in CENCALC to help overcome this
34 limitation.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52 *Method-1: Mutual Information Expansion*

53
54
55 The entropy estimation by MIE is computed through the generalized mutual information
56 functions.⁹ For two variables A and B , representing in this case the conformational state of two
57
58
59
60

torsion angles, their mutual information, $I(A, B)$, is defined in terms of entropies as $I(A, B) = S(A) + S(B) - S(A, B)$, where $S(A)$ and $S(B)$ are the entropies of A and B , respectively, and $S(A, B)$ is the joint entropy of AB . In general, the mutual information shared among k variables is expressed as

$$I_k(A_1, \dots, A_k) = \sum_{m=1}^k (-1)^{m+1} \sum_{\substack{J \subset \{A_1, \dots, A_k\} \\ |J|=m}} S(J) \quad (2)$$

For a given value of m , the inner sum in the latter expression runs over all possible subsets of $I = \{A_1, \dots, A_k\}$ with m elements. On the basis of the mutual information functions, the n -order MIE leads to the following approximation to the full entropy:

$$S^{(n)}(A) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{I \subset \{A_1, \dots, A_M\} \\ |I|=k}} I_k(I) \quad (3)$$

For example, the MIE expression at second order corresponds to the well-known pairwise approximation:

$$S^{(2)}(A) = \sum_{i=1}^M S(A_i) + \sum_{i < j}^M \{S(A_i, A_j) - S(A_i) - S(A_j)\}, \quad (4)$$

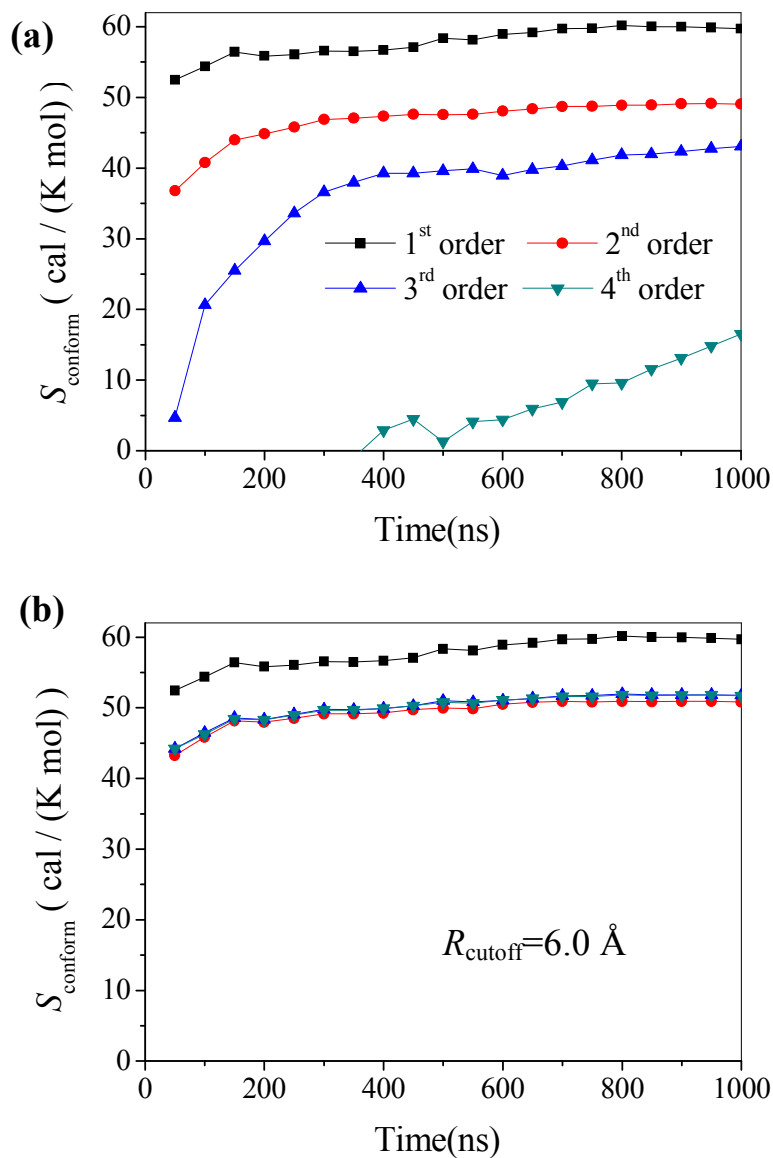
In principle, the implementation of the original MIE expression, Eq. (3), may result a in computational bottleneck because the entropy of any subset $\{A_i, A_j, \dots\}$ of A with cardinality less than n must be either recomputed or stored and accessed repeatedly during the calculation of the higher order entropy terms. In previous work,²¹ we have shown that Eq. (3) admits the following alternative expression that removes all redundancy in the calculation of the MIE terms,

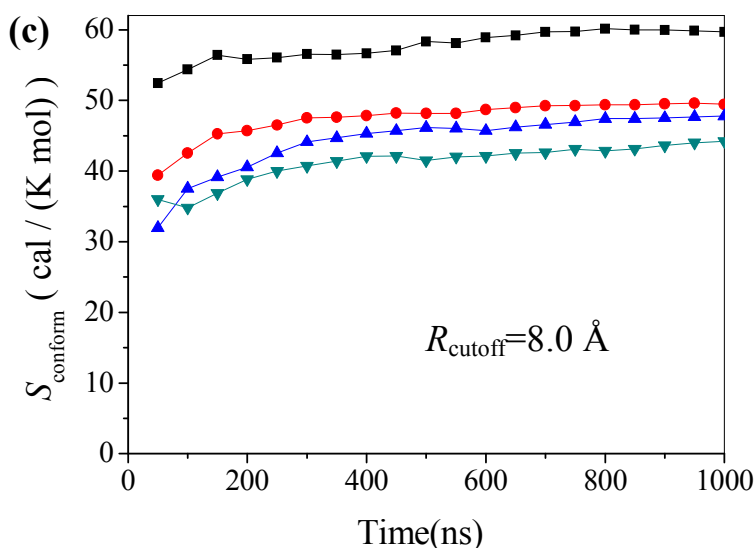
$$S^{(n)}(\mathbf{A}) = \sum_{k=1}^n \left[\sum_{i=0}^{n-k} (-1)^i \binom{M-k}{i} \right] \sum_{\substack{I \subset \{A_1, \dots, A_M\} \\ |I|=k}} S(I) \quad (5)$$

and that is implemented in the CENCALC program.

Figure 4a displays a series of entropy plots that were obtained by applying the MIE method. All the rotatable bonds of the **GNR** peptide were taken into account ($M=44$) and various expansion orders ($n=1-4$) were considered in order to include correlation effects among all the torsion angles regardless of their relative separation. First, we note that the marginal entropy, that is, the first order entropy that corresponds to the sum of the independent entropies of the torsion angles, $\sum_i^M S(A_i)$, is reasonably well converged after having averaged over $\sim 10^6$ configurations, leading to a limiting value of ~ 60 cal/(K mol) (the convergence gradient of $\sum_i^M S(A_i)$ is lower than 3×10^{-3} cal/(K mol ns)). Similarly, the $S^{(2)}(\mathbf{A})$ plot reaches a rather flat plateau showing that entropy reduction due to pair correlation effects amounts to ~ 9 cal/(K mol). However, it is also clear that the $S^{(3)}(\mathbf{A})$ entropy plot in Figure 4a and, especially, the fourth-order approximation, are quite far from being converged with respect to the simulation time. Thus, we did not compute higher order terms because they would result in strongly biased entropies due to sampling limitations, apart from being computationally very expensive due to the combinatorial explosion implicit in the conventional MIE method.

Figure 4 Convergence plots of the GNR peptide as obtained with the MIE method at various orders: (a) using no cutoff and Eq.(5); (b) and (c) using cutoff values of 6.0 and 8.0 Å, respectively, and Eq. (6).





24 Besides the issue of convergence with respect to the simulation time, the conventional
25 MIE calculations on medium-sized molecules suffer from convergence problems with reference
26 to the expansion order n that is required to capture correlation effects. Only in the case of
27 relatively small molecules (alkanes, dipeptides,...), we have observed MIE-order convergence.²¹
28 Moreover, even in the case of perfectly time-converged calculations, truncation errors would be
29 present and they would not necessarily decrease with the expansion order, that is, high order
30 calculations do not always result in better entropy estimations. For example, from the
31 calculations on the **GNR** decapeptide summarized in Figure 4a, it is still uncertain whether or not
32 the 2nd or the 3rd-order MIE calculations may lead to a reasonable approximation to the exact
33 value.
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 In many potential applications of conformational entropy calculations (*e.g.*, entropy
49 changes of ligand molecules upon binding to a receptor), it could be convenient to focus on those
50 short- and medium-range correlations that are presumably more important. By discarding the
51 inclusion of long-range effects in the entropy calculations, better convergence properties should
52 also be expected. This can be done by applying a well known and widely used approximation in
53
54
55
56
57
58
59
60

Computational Chemistry, that is, the use of a distance-based cutoff criterion R . Thus, CENCALC offers the possibility of performing MIE calculations with a predefined cutoff value. When doing so, the inner sum in Eq. (3) is restricted to those subsets of A in which the mean distance d_{ij} between any pair of elements A_i and A_j is less than R , resulting in the following R -dependant MIE expression:

$$S_R^{(n)}(\mathbf{A}) = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{J \in C(R) \\ |J|=k}}^M I_k(J) \quad (6)$$

where $C(R) := \left\{ I \subset \{A_1, \dots, A_M\} \mid \max_{A_i, A_j \in I} \{d(A_i, A_j)\} < R \right\}$ is the class of subsets that meets the R cutoff criterion. This new expression maintains the invariance of the total entropy under any $A_i \leftrightarrow A_j$ permutation and contains a large degree of redundancy likewise the original MIE equation.

We see in Figure 4b and Figure 4c that cutoff-restricted MIE calculations result in entropy plots having much improved convergence properties with respect to the expansion order. For a relatively small cutoff value of $R=6.0 \text{ \AA}$, we find that the $S_{conform}$ of **GNR** at the 2nd, 3rd and 4th orders all give nearly coincidental plots that reach a stable plateau along the simulation time. Of course this is the consequence of having ignored a significant fraction of correlation among the torsional motions of **GNR**. When R is increased up to 8 \AA (see Figure 4c), the correlated entropy curves keep an acceptable convergence behaviour, but they tend now to different limiting values lying in the $40\text{-}50 \text{ cal}/(K \text{ mol})$ interval. Hence, the cutoff-dependant MIE method can help obtaining entropy curves with better convergence properties at the cost of introducing a new empirical parameter that needs to be assessed (*i.e.*, which is the best cutoff?). For intermediate cutoff values, it does not remove order truncation errors either. Furthermore, it turns out that the

cutoff-restricted MIE calculations are still computationally very expensive at high orders (see below). Fortunately, these limitations can be substantially mitigated by using other methods that are also available in CENCALC.

Method-2: Approximate Mutual Information Expansion (AMIE)

To partially mitigate the efficiency drawbacks of the conventional and cutoff-restricted MIE calculations, the CENCALC software implements an *approximate* MIE (AMIE) method.²⁶ The AMIE method is based on the construction of a non-redundant neighbour list $L = \{L_1, L_2, \dots, L_M\}$, which, in turn, is a list of M sublists L_i . The first element of each L_i is A_i and the rest of members are selected among the neighbours A_j of A_i so that $j > i$ and $d_{ij} < R$ (see reference²⁶ for further details). This approach has allowed us to propose the following expression

$$S_L^{(n)} = \sum_{k=1}^n (-1)^{k-1} \sum_{i=1}^M \sum_{\substack{J \subset (L_i - \{A_i\}) \\ |J|=k-1}} I_k(A_i, J), \quad (7)$$

as an approximation to Eq. (6). By carrying out some algebraic manipulation, the latter expression can be transformed into a computationally efficient form:

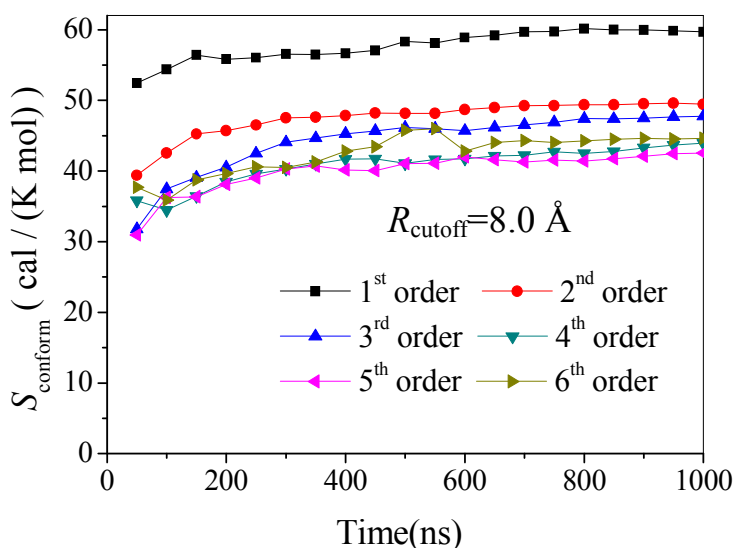
$$S_L^{(n)} = \sum_{i=1}^M \sum_{k=1}^n \left[\sum_{j=0}^{n-k} (-1)^j \binom{|L_i| - k}{j} \right] \sum_{\substack{J \subset (L_i - \{A_i\}) \\ |J|=k-1}} (S(A_i \cup J) - S(J)), \quad (8)$$

which corresponds to the AMIE method as implemented in CENCALC.

In a previous work,²⁶ we have shown that the main difference between the AMIE and MIE formulations is that AMIE is not strictly invariant under $A_i \leftrightarrow A_j$ exchanges. Nevertheless, the

AMIE implementation gives a significant speedup and yields results that are numerically very close to those provided by the original MIE formulation (e.g., the largest difference in the limiting entropy values is only 0.05 cal/(K mol) for **GNR**). As shown in Figure 5, the AMIE calculations allow us to estimate various entropy terms up to 6th-order using a 8.0 Å cutoff. The AMIE $S_{conform}$ estimations for **GNR** derived from the limiting values at the 4th, 5th and 6th-orders are 44, 43 and 45 cal/(mol K), which correspond to an uncertainty of only 0.6 kcal/mol in terms of free energy at 300 K.

Figure 5 Convergence plots of the **GNR** peptide as obtained with the AMIE method at various orders with a 8.0 Å cutoff and using Eq. (8).



Method-3: Multibody Local Approximation (MLA)

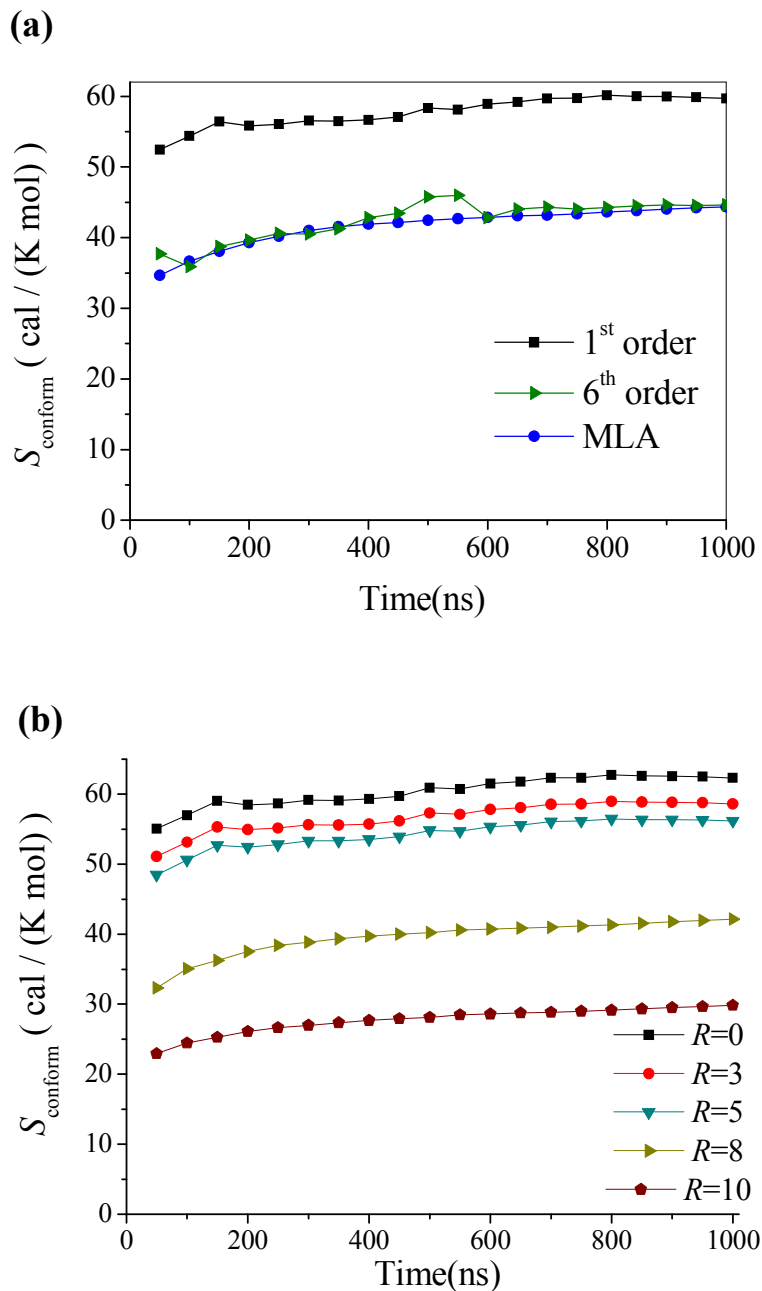
To make sure that order convergence is achieved in the cutoff-dependant entropy calculations, it is clear that a better strategy would be to consider all the n -order effects that are allowed by the predefined cutoff. This task, which might seem *a priori* computationally intractable in the general case, can be fulfilled by the Multibody Local Approximation (MLA) implemented in CENCALC for large and medium-sized molecules. The MLA expression used by CENCALC has been

1
2
3 derived in our previous work²⁶ starting from the AMIE expression. By summing first over the
4 number of lists $i=1, \dots, M$ and enabling that the expansion order n can take its maximum value
5
6 within L_i , a largely simplified entropy expression is obtained:
7
8
9

$$S_L = \sum_{i=1}^M [S(L_i) - S(L_i - \{A_i\})] \quad (9)$$

10
11
12
13
14
15
16
17 where $S_L(A)$ is the Multibody Local Approximation for the calculation of $S(A)$ that is
18
19 effectively independent of any expansion order.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 6 (a) Convergence plots of the GNR peptide as obtained with the 6th-order AMIE and MLA methods and using a 8.0 Å cutoff. (b) MLA entropy plots at different cutoff values (R in Å)



In

1
2
3
4
5
6
7 Figure 6(a) we compare the **GNR** entropy plots computed with the 6th-order AMIE and MLA
8
9 methods using the same cutoff (8 Å). By construction, the MLA estimation is identical to a
10
11 maximum order AMIE calculation and, therefore,
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Figure 6 confirms that the 6th-order expansion using MIE is a reliable approximation to the total
8 entropy with $R=8.0$ Å. However, the important message here is that the MLA method largely
9 outperforms the AMIE calculations, not only because MLA includes implicitly all order effects
10 within the predefined cutoff, but also for its large speedup (see below). Therefore, we conclude
11 that, for a given cutoff, the MLA implementation solves the problem of order convergence and
12 performs much more efficiently than the AMIE protocol.
13
14
15
16
17
18
19

20
21 Of course, the limiting value of the MLA entropy depends significantly on the cutoff
22 value R as clearly shown in
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Figure 6(b). For example, after 1.0 μ s of simulation time, the estimated $S_L(A)$ values for GNR
8
9 are 59, 56, 42 and 30 cal/(K mol) for $R=3, 5, 8$ and 10 \AA , respectively. The corresponding
10
11 entropy plots have also similar convergence properties as they result in similar convergence
12
13 numerical gradients of $\pm 3 \times 10^{-3}$ cal/(K mol ns). Regardless of the increasing computational cost
14
15 of the MLA calculations with R , one could expect that lower $S_L(A)$ values should be obtained
16
17 at larger cutoff values because of the capture of *presumably* more *physical* correlation. However,
18
19 the use of larger cutoffs (if computationally affordable) does not necessarily result in better
20
21 entropy estimations because entropy estimations derived from a finite amount of data tend to be
22
23 considerably biased⁴⁰, resulting in a systematic underestimation of the true entropy, or
24
25 equivalently, an overestimation of the correlation effects. In the context of the MLA $S_{conform}$
26
27 calculations, this means that a larger cutoff augments the sample space (*i.e.*, the number of
28
29 possible conformational states included in the lists L_i and $L_i - \{A_i\}$ in Eq.(9)) whereas the
30
31 available amount of data (*i.e.*, the number of MD snapshots) remains fixed. In other words, as the
32
33 cutoff increases, the estimated $S_{conform}$ decreases due to the capture of more physical correlation,
34
35 but also due to the growth of negative bias or false correlation.
36
37
38
39
40
41
42
43
44
45
46
47

48 *Method-4: Correlation-corrected MLA (CC-MLA)*

49
50

51 Clearly, the proper use of the MLA approach is linked to the problem of determining the best
52
53 cutoff R for estimating the conformational entropy from a given amount of MD sampling. To
54
55 help solving such problem, we have developed the so-called correlation-corrected MLA entropy
56
57 estimator (CC-MLA).²⁶ The derivation of the CC-MLA estimator assumes that all the entropy
58
59
60

bias in the calculation is entirely due to false correlation effects and, consequently, its practical application is valid only if the marginal entropies of all the rotatable bonds are reasonably converged within the simulation time. Under this assumption, the original MIE expression in Eq. (3) can be rewritten for $n=M$ in order to express the *exact* entropy as follows:

$$S_{exact}(A) = \sum_{A_i \in A} S(A_i) + \sum_{k=2}^M (-1)^{k-1} \sum_{\substack{J \subset A \\ |J|=k}} \left\{ E[\hat{I}_k(J)] - \text{Bias}[\hat{I}_k(J)] \right\} \quad (10)$$

where $S(A_i)$ are the presumably *exact* values of the marginal entropies, $E[\hat{I}_k(J)]$ is the expected value of the sample mutual information $\hat{I}_k(J)$ shared among the components of J , and $\text{Bias}[\hat{I}_k(J)]$ is the bias of $\hat{I}_k(J)$.

After some statistical approximations and algebraic manipulations,²⁶ the latter expression Eq. (10) can be transformed into the following asymptotically unbiased estimator

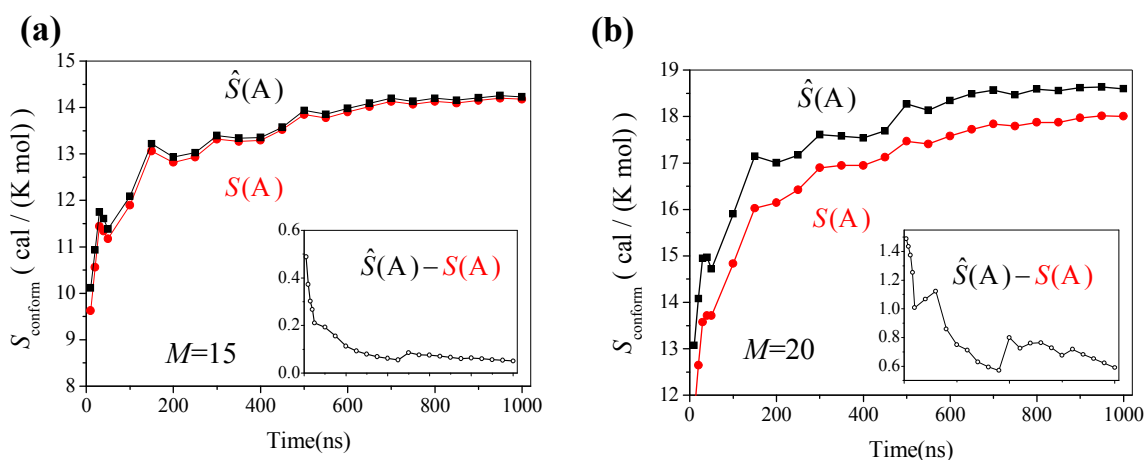
$$\hat{S}(A) = \sum_{A_i \in A} S(A_i) + [S(A) - S'(A)] \quad (11)$$

where $S(A)$ is the sample entropy, that is, the entropy computed from the sample without any further consideration or approximation, and $S'(A)$ is the sample entropy computed for the M -dimensional random variable X' representing the conformational state of A after having independently and randomly reordered the outcomes of at least $M-1$ of its components.²⁶

It may be interesting to compare the new estimator $\hat{S}(A)$ with the classical one $S(A)$. To this end and for the sake of simplicity, we consider two subsets of the **GNR** torsion angles composed by the first 15 and 20 torsion angles in the topology file, respectively, for which *exact*

1
2
3 sample entropy calculations using the Shannon expression (Eq. (1)) can be made. As shown in
4
5 Figure 7, the discrepancy between the two estimators is greater for the larger set because the
6
7 bigger conformational space, the more biased entropy estimates. On the other hand, it can be
8
9 expected that the $\hat{S}(A) - S(A)$ difference would decrease with a longer simulation time (*i.e.*, a
10
11 larger sample size).
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 7. Comparison between the corrected $\hat{S}(A)$ (Eq. (11)) and uncorrected $S(A)$ estimators. (a): Using the first 15 torsional angles from the GNR topology file. (b): Using 20 torsional angles.



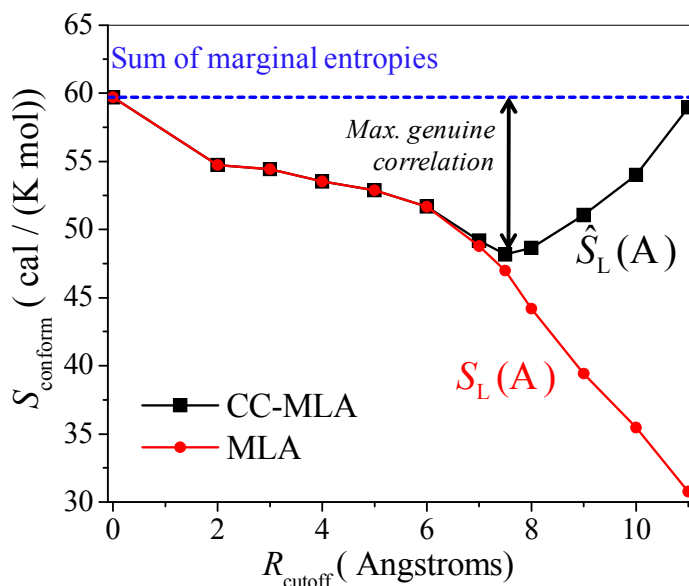
The new estimator $\hat{S}(A)$ can also be used directly with MLA to obtain the corresponding

CC-MLA method:

$$\hat{S}_L(A) = \sum_{A_i \in A} S(A_i) + [S_L(A) - S'_L(A)]. \quad (12)$$

Perhaps the most interesting feature of the CC-MLA estimator is that it allows us to select the *best* cutoff for every particular case of study, retrieving thus the genuine correlation from a given amount of MD sampling.²⁶ As shown in Figure 8, the plot of $\hat{S}_L(A)$ vs. R has a global minimum, which has been shown to correspond to the optimal cutoff and the lowest upper bound of the CC-MLA entropy estimation. In general, the optimal cutoff will depend on the nature of the molecular system and the available amount of sampling. The CC-MLA method is the default method in CENCALC and the recommended one for medium-sized and large molecules.

Figure 8. Comparison between the MLA and CC-MLA entropies for the 1.0 μ s trajectory of the GNR peptide at different cutoff values.



Computational efficiency of the entropy methods implemented in CENCALC

Clearly, the applicability of the MIE, AMIE, MLA and CC-MLA methods will depend on different factors like the size and flexibility of the molecular system, the length of the simulations, whether or not absolute entropies are required, the cancellation of errors in the computation of relative entropies, etc. All these factors influence directly (or through the assessment of) the convergence properties of the entropy methods with reference to the amount of sampling and/or the expansion order in the MIE-like methods. However, all the potential CENCALC users should also be aware of the performance of the entropy methods in terms of their computational efficiency, the *local* methods, MLA and CC-MLA, having a great advantage in this respect over the MIE-like expansions.

1
2
3 In
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 9 the computational cost of the different entropy methods is represented as a function of the number of snapshots (N) of the **GNR** peptide employed in the calculations. More specifically, entropies were computed using a 8.0 Å cutoff and a 6th-order expansion with the AMIE and MIE methods. The CPU time data are also presented in Table S1 together with the corresponding entropy plots in Figure S1 in the Supporting Information. As the 6th-order MIE/AMIE and MLA methods give very similar entropy values (~44 cal/(K mol); see also Figure 5 and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 6), a fair comparison among their CPU time consumptions can be made. The CC-MLA estimation, which should remove some negative bias, gives a limiting value of entropy that is about 4 cal/(K mol) more positive, but the corresponding CPU times are directly comparable to those of the other methods.

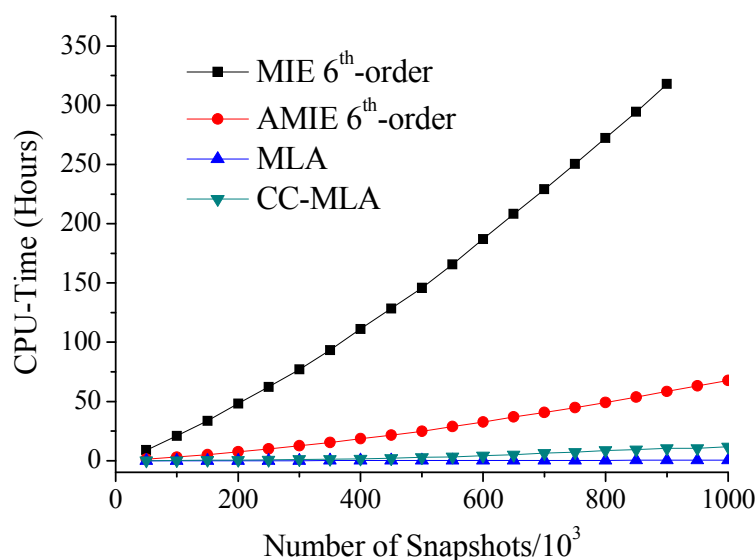
For the four entropy methods, we found that their computational cost scales nearly linear with N (although the data fit slightly better to a second order polynomial). However, the slope of the curves in

Figure 9 changes dramatically depending on the entropy method. For example, the MIE method is computationally much more expensive than either AMIE or MLA. This is basically due to the fact that the CENCALC implementation of the cutoff-based MIE method requires the evaluation of thousands and millions of entropy terms that are computed repeatedly on-the-fly as they are needed to obtain high order entropy contributions. Even though a more efficient programming of the MIE method could be possible by storing intermediate data in the rapid access memory, it can be safely expected that the intrinsic advantage of the AMIE and MLA methods, which provide $\sim 6\times$ and $\sim 10^3\times$ speedups in the computation of the conformational entropies, respectively, would be essentially preserved. Most remarkably, the rather low computational cost of the MLA method demonstrates its adequacy for studying large systems. However, by keeping in mind the convenience of optimizing the R cutoff, it must be noted that the default CC-MLA method is more expensive than MLA, as shown by our test calculations in which CC-MLA performs worse than MLA by a factor of ~ 20 . The main reason for this loss of performance can be traced back to the independent random sorting of each of the outcomes of the torsion angles (*i.e.*, the columns in the *MATRIX.dat* input file) that is characteristic of the CC-MLA estimator. This operation

destroys the correlation among the conformational states (*i.e.*, the rows in *MATRIX.dat*) what, in turn, has a negative impact in terms of CPU time in different parts of the *cencalc_omp.f90* code as, for example, in the match search routines that compare a given conformational state with all the former ones. Nevertheless, in spite of its relatively large overhead cost, there is no doubt that, globally, CC-MLA has by far a much better behaviour than AMIE or MIE (see

Figure 9) and provides a reasonable trade-off between cost and statistical accuracy.

Figure 9 Total CPU time (in hours) vs the number of snapshots processed by the conformational entropy calculations using the various methods implemented in CENCALC with a 8.0 Å cutoff and using the coordinates of the 1.0 μs GNR simulation. CPU time data correspond to a single-core of one Intel Xeon X545 0 processor (3.0 Ghz, 6Mb cache).



TEST APPLICATIONS

1
2
3 In previous works, we have carried out conformational entropy calculations in order to assess the
4 role of entropy in the absolute and relative stability of collagen model peptides²³ as well as in the
5 binding of small peptides to the active site of the MMP-2 matrix metalloprotease.³⁶ In these
6 preliminary calculations, we employed the conventional MIE approach, which was then the only
7 method implemented in CENCALC, and the dimensionality of the problems had to be reduced by
8 truncating the MIE expansion at fourth order and analyzing only the backbone conformational
9 variability of the peptide molecules. The obtained $S_{conform}$ values were used to complement
10 approximate MM-PB free energies that included the average RRHO entropies. More recently, we
11 have shown that the combination of \bar{S}_{RRHO} with the $S_{conform}$ values obtained from the latest version
12 of CENCALC may have a broader applicability as suggested by test calculations on systems
13 ranging from hydrocarbon molecules in the gas-phase to a polypeptide molecule in aqueous
14 solution.²¹ The sophistication, computational efficiency and diversity of the entropy methods
15 currently available in CENCALC should allow potential users to perform more thorough analyses
16 of conformational entropy in complex molecular systems. To further illustrate this point, herein
17 we briefly comment on two particular cases of study that are representatives of the kind of
18 problems to which the CENCALC entropy calculations could be particularly amenable.
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 *Dynamics of the isolated GNR peptide in solution*

46
47
48 Whenever the entropy of subgroups of torsion angles is of interest, the use of the CENCALC
49 software is straightforward. Only those torsion angles of interest are then included in the
50 computation. In the case of the **GNR** peptide, for example, the conformational entropy
51 contributions arising from rotatable bonds involving only either backbone or side-chain atoms
52 can thereby be analyzed separately. Moreover, segregation of the backbone and side chain
53
54
55
56
57
58
59
60

1
2
3 contributions allows us to analyze the origin of the notable reduction in $S_{conform}$ (~40-50%) due to
4
5 correlation effects with respect to the sum of the marginal entropies, which assumes independent
6
7 torsional motions (see Figures 4-6). For this purpose, examining the results of the MIE-like
8
9 methods at various orders seems particularly appropriate because these methods can reveal the
10
11 extent and complexity of correlation effects provided that sufficiently converged data could be
12
13 obtained. Thus,
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Table 1 collects the limiting values of the entropy contributions from all the torsion angles
8 ($M=44$), the backbone torsion angles ($M=16$), and those in the residue side chains ($M=28$), and
9 that were computed at the 1st (marginal entropies), 2nd (pair correlation) and 6th (high order
10 correlation effects) expansion orders using the AMIE method with an optimal $R=7.5$ Å cutoff as
11 suggested by the previous CC-MLA entropy estimation. These various contributions to the
12 conformational entropy level off at the end of the trajectory, indicating a quite stable convergence
13 (see Figure S2 in the Supporting Information).
14
15
16
17
18
19
20
21
22
23

24 Comparing the limiting values of the entropy of the whole peptide, the backbone, and the
25 side chains, it turns out that the backbone contributes less entropy than do the side-chains both in
26 absolute and relative terms (see Table 1). For example, the $S_{conform}^{(1)}$ values per torsion angle
27 amounts to 0.862 and 1.761 cal/(K mol) for the backbone ($M= 16$) and side chain ($M=28$)
28 torsions, respectively, which in terms of percentage of total entropy translate into 23 and 77 % .
29 This larger share of the entropy in the side-chains can be attributed to their greater flexibility
30 while the main chain of **GNR** tends to adopt a helical structure as shown in Figure 3. Of more
31 interest can be the analysis of the correlation effects as measured by the $S_{conform}^{(2)}$ and $S_{conform}^{(6)}$
32 values
33 collected
34 in
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Table 1. Thus, pair correlation has only a minor role in backbone motions (1.5 cal/(K mol)), but
8
9 reduces significantly the entropy of the side chain motions (10.9 cal/(K mol)). Similarly, high
10
11 order correlation effects within the predefined cutoff hardly affect backbone entropy with respect
12
13 to either the first or second order entropies, but they clearly influence the $S_{conform}$ of the **GNR** side
14
15 chains as the corresponding $S_{conform}^{(6)}$ value lies 3.8 cal/(K mol) above $S_{conform}^{(2)}$ (*i.e.*, high order
16
17 contributions do not necessarily lead to lower entropy values). Concerning the correlation effects
18
19 between backbone and side-chain torsions, we observe a notable reduction in the total entropy,
20
21 that is, $\Delta S_{conform}^{(6)} = S_{conform}^{(6),total} - S_{conform}^{(6),backbone} - S_{conform}^{(6),side\ chain} = -6.3$ cal/(K mol). Note also that this effect
22
23 can only be taken into account through high order contributions connecting torsion angles of
24
25 different residues. This confirms that the coupling between the backbone and side chain motions
26
27 is an important contribution to correlation effects in the conformational entropy of the **GNR**
28
29 peptide and that its explicit consideration requires the use of high order expansion entropy
30
31 methods and a considerable sampling effort.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1 Number of rotatable bonds (M) and limiting values of the n -order AMIE ($n=1, 2$ and 6) and CC-MLA (with optimal cutoff) conformational entropies (in cal/(K mol)) for the **GNR** peptide after 1.0 μ s (unbound form)/ 0.4 μ s (MMP-7 bound form) of simulation time. Percentages of backbone and side-chains contributions (in parentheses) and entropy values per rotatable bond (in squared brackets) are also indicated.

	Total	Backbone	Side-chains
M	44	16 (36%)	28 (64%)
Unbound form			
AMIE $S_{conform}^{(1)}$	59.7 [1.36]	13.8 (23%) [0.862]	47.9 (77%) [1.71]
AMIE $S_{conform}^{(2)}$	49.6 [1.13]	12.3 (25%) [0.768]	37.0 (75%) [1.32]
AMIE $S_{conform}^{(6)}$	46.9 [1.06]	12.4 (26%) [0.775]	40.8 (74%) [1.46]
CC-MLA	48.2 [1.09]	12.4 (26%) [0.775]	41.0 (74%) [1.46]
MMP7-bound form			
M	33	6	27
AMIE $S_{conform}^{(1)}$	42.3 [1.28]	6.29 (15%) [1.05]	35.9 (85%) [1.33]
AMIE $S_{conform}^{(2)}$	34.5 [1.04]	5.89 (18%) [0.982]	30.9 (82%) [1.14]
AMIE $S_{conform}^{(6)}$	31.0 [0.94]	5.89 (19%) [0.982]	27.7 (81%) [1.02]
CC-MLA	34.8 [1.05]	5.89 (17%) [0.982]	29.7 (83%) [1.10]

Conformational entropy change upon binding of GRN to MMP-7

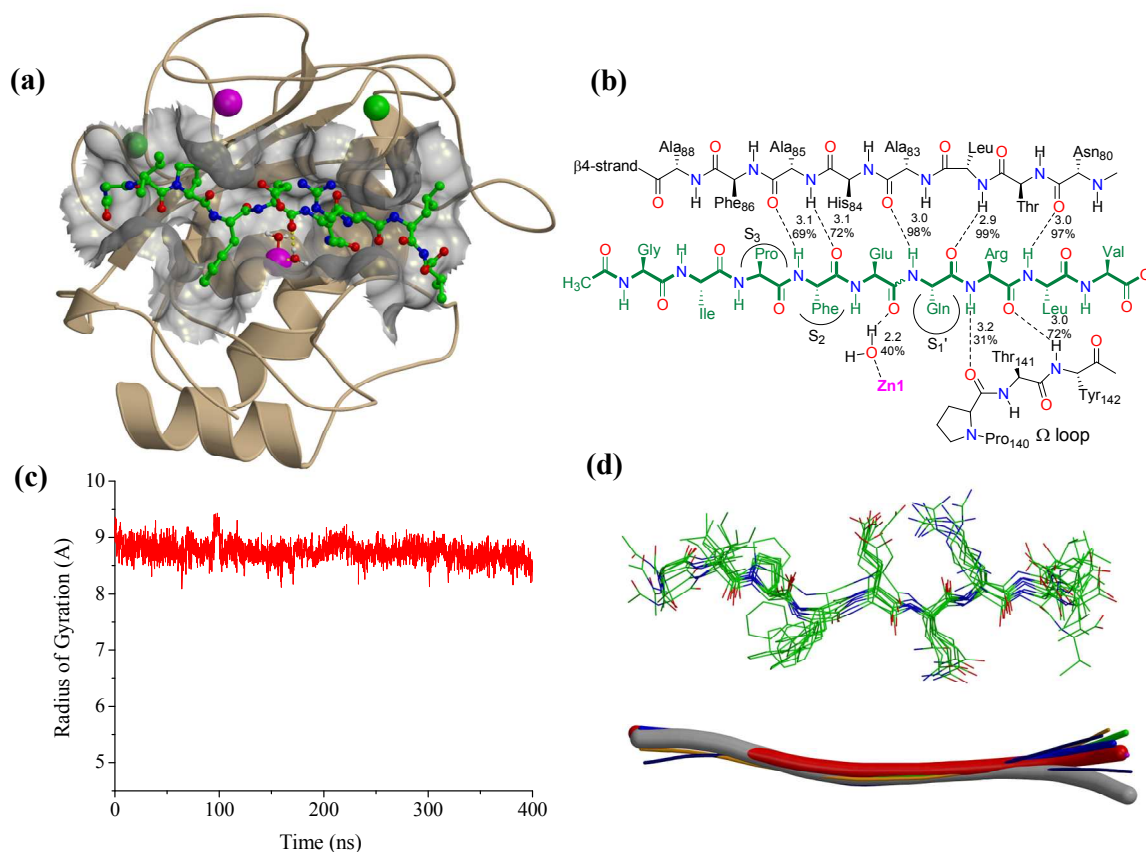
In computer assisted drug design, free energy calculations play a key role in unraveling the thermodynamical forces controlling biomolecular association processes and/or making predictions about ligand binding.^{2,41} Given that $\Delta G_{binding}$ comprises both enthalpic and entropic contributions, understanding the origin of these contributions is equally important to find out the actual binding determinants explaining the activity of substrate or inhibitor molecules. In the case of binding entropy, it has been shown that the two major contributions are the change in conformational entropy and the change in solvation entropy.¹⁶ The latter contribution, which is taken into account implicitly by solvent continuum models as those used in approximate free energy methods or in physically-based scoring functions, is particularly difficult to extract from computer simulations with explicit solvent as it requires the use of thermodynamic cycles and free energy perturbation (FEP) calculations.¹⁶ A similar strategy is followed by the restraint release method to compute the change in the configurational entropy upon ligand binding that includes also the change in conformational entropy.¹⁵⁻¹⁶ However, direct entropy methods as those implemented in CENCALC offer an alternative procedure, potentially simpler, to estimate the conformational entropy contributions to binding entropy that, in turn, can be utilized within the context of less rigorous theoretical methods^{24,42-43} (in these approximated methods, it is generally assumed that free energies or entropy contributions can be attributed to particular degrees of freedom and/or physical interactions).

To further illustrate the utility of the CENCALC $S_{conform}$ calculations in the analysis of binding processes, we simulated the complex formed between **GNR** and the MMP-7 enzyme. This enzyme is secreted by tumour cells and has been validated as an anticancer drug target.²⁷ As an MMP-7 substrate, **GNR** has a specificity kinetic constant of $425 \text{ M}^{-1}\text{s}^{-1}$ that results from

1
2
3 moderate binding ($K_M=3.7$ mM) and catalytic efficiency ($k_{cat}=1.6$ s⁻¹) as compared with other
4
5 peptide sequences.²⁵
6
7

8
9 Figure 10 shows the positioning of **GNR** in the MMP-7 active site and the main H-bond
10 contacts that stabilize **GNR** in an extended conformation characterized by an r_{gyr} value of 8.7 ± 0.2
11 Å. A total of seven H-bond contacts (67-100% of occupancy) connect the important backbone
12
13 positions in the β 4-strand and the Ω -loop of the enzyme with the corresponding backbone amide
14
15 groups of **GNR** (Ala₈₅-O \cdots HN-Phe(P_2), Ala₈₅-NH \cdots O-Phe(P_2), Ala₈₃-O \cdots HN-Gln(P_1'), Tyr₁₄₂-
16
17 NH \cdots O-Arg(P_2'), ..., see Figure 10). Hydrophobic contacts also contribute to anchor the **GNR**
18
19 substrate within the MMP-7 active site as the Pro(P_3) and the Phe(P_2) residues remain
20
21 accommodated in the so-called S_3 and S_2 hydrophobic pockets all along the simulation time. In
22
23 addition, the carbonyl group of the scissile Glu(P_1)~Gln(P_1') peptide bond binds the catalytic
24
25 zinc ion either directly (during the first 160 ns of simulation time) or through a Zn \cdots (H₂O) \cdots O=C
26
27 water bridge that is formed upon a conformational transition of the **GNR** backbone chain and
28
29 remained stable for the rest of the simulation.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 10 (a) Ribbon models and molecular surface representations of a MD snapshot from the **MMP-7/GNR** trajectory (the **GNR** peptide is depicted in ball-and-sticks with carbon atoms shown in green). (b) Schematic representation of the main enzyme/peptide binding determinants (average distances are in Å). (c) Radius of gyration along the MD simulation for the isolated **GNR** peptide. (d) Superposition of the representative structures accounting for 90% of backbone variability according to clustering analyses, both in wireframe and ribbon model representation. Thickness of the ribbon models corresponds to the number of snapshots represented by each model.



1
2
3 Comparison of the results of the clustering analyses for the unbound and MMP-7 bound
4 forms of the **GNR** peptide (see Figure 3 and Figure 10), clearly shows that binding to MMP-7
5
6 reduces significantly the dynamical flexibility of the backbone and side chains of **GNR**. In fact,
7
8 the post-processing of the **MMP-7/GNR** trajectory by the *cencalc_prep.f90* code reveals that
9
10 only six torsion angles of the **GNR** main chain exhibit conformational variability, that is, 10
11 backbone torsion angles become frozen upon substrate binding. This loss of flexibility is also
12
13 reflected in the faster time convergence of the S_{conform} calculations (0.4 μs) as well as in the larger
14
15 value of the optimal cutoff value for the CC-MLA entropy estimation (16 \AA), the CC-MLA
16
17 entropy vs R plot being actually quite flat in the 8-20 \AA interval (see Figures S3 and S4 in the
18
19 Supporting Information). In terms of the limiting values of entropy, it turns out that binding
20
21 reduces by ~50% and ~25% the backbone and side chain conformational entropies, respectively,
22
23 according to either 6th-order or CC-MLA methods (see
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Table 1). Moreover, the formation of the enzyme-peptide interactions also disrupts the highly
8
9 correlated intramolecular motions of the unbound **GNR** peptide. Thus, only pair correlation
10
11 effects seem to be important when **GNR** lies in the MMP-7 active site given that the $S_{conform}^{(2)}$ and
12
13
14 $S_{conform}^{(6)}$ values are now very similar to each other (see
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 Table 1). Globally, our best estimation for the CC-MLA $S_{conform}$ of the **GNR** peptide bound to
8
9 MMP-7 is ~ 13 cal/(K mol) below that for the unbound state, what results in a significant free
10
11 energy penalty of ~ 4 kcal/mol at 300 K. If independent torsions are assumed, then the resulting
12
13 conformational penalty is slightly larger (~ 5 kcal/mol). It may also be interesting to note that
14
15 although other entropic and free energy contributions should be considered to obtain a complete
16
17 thermodynamic picture, the $-T \Delta S_{conform}^{binding}$ term estimated for the substrate molecule would be most
18
19 likely the major conformational penalty to **GNR** binding given that no significant changes in the
20
21 active site region of the MMP-7 enzyme occur upon peptide binding.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

SUMMARY AND CONCLUSIONS

In this work we have presented the CENCALC software for the estimation of the conformational entropy of a flexible single molecule. This entropy contribution stems from the partitioning of the intramolecular entropy into vibrational and conformational contributions as originally proposed by Karplus and co-workers, and that we have analyzed in detail in a previous work.²¹ The current version of CENCALC includes an auxiliary script for interfacing with the trajectory and topology files from MD simulations performed with the Amber package. However, users of CENCALC could easily adapt it to process data generated with other MD software or Monte Carlo simulations.

Prior to entropy calculations, CENCALC discretizes the evolution of the torsional degrees of freedom for the molecule of interest. This preliminary transformation greatly facilitates the development of computationally efficient entropy methods based on the mutual information expansion that can capture high order correlation effects contributing to conformational entropy. In this respect, users of CENCALC can access to an array of techniques (MIE, AMIE, MLA and CC-MLA) that can be combined with an interatomic distance-based cutoff in order to obtain converged entropy estimations for many potential systems of interest. In any case, it is clear that intensive sampling in the range of hundreds of ns may be required in order to converge the conformational entropy calculations of flexible molecules. Fortunately, these considerable sampling requirements are becoming routinely accessible thanks to impressive breakthroughs in software and hardware technologies.⁴⁴

While the MIE-like methods are probably best suited for the study of relatively small molecules, the default method in CENCALC, CC-MLA, should be preferable for medium-sized

1
2
3 and large molecules provided that reasonably converged values of marginal entropies are
4
5 obtained. For these systems, the removal of the negative entropy bias due to sampling limitations
6
7 can be done by selecting the R cutoff value that minimizes the corresponding CC-MLA
8
9 estimation. In any case, a careful monitoring and assessment of the convergence properties would
10
11 be a prerequisite for obtaining meaningful results. Furthermore, the judicious choice of subsets of
12
13 torsion angles for conformational entropy calculations and/or the computation of relative entropy
14
15 changes rather than absolute ones, will help make CENCALC a useful computational tool for
16
17 obtaining new insight into the dynamical properties of flexible molecules as well as for
18
19 complementing the results of approximate free energy calculations.
20
21
22
23
24
25
26
27

28 **ACKNOWLEDGMENT**

29
30
31
32 This research was supported by the following grants: FICyT (Asturias, Spain) IB05-076 and
33
34 MEC (Spain) CTQ2007-63266. E.S. thanks MEC for his FPU contract (AP2005-4882).
35
36
37
38
39
40

41 **SUPPORTING INFORMATION**

42
43
44 Table S1 and Figure S1 summarizing the results of the computational cost of entropy
45
46 calculations. Figure S2 showing AMIE and CC-MLA entropy plots for the **GNR** peptide
47
48 discriminating between total, backbone, and side chain contributions. Figure S3 comparing
49
50 between MLA and CC-MLA entropies of the MMP-7 bound **GNR** peptide at different cutoff
51
52 values. Figure S2 showing AMIE and CC-MLA entropy plots for the MMP-7 bound form of
53
54 **GNR** discriminating between total, backbone, and side chain contributions. CENCALC user's
55
56
57
58
59
60

1
2
3 manual in PDF format. A Zip file containing the CENCALC software and numerical examples.
4
5

6 This material is available free of charge via the Internet at <http://pubs.acs.org>.
7
8
9

10 11 12 REFERENCES 13 14 15 16 17

18 (1) Baron, R.; Gunsteren, W. F. v.; Hünenberger, P. H. Estimating the Configurational
19 Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to
20 the Quasi-Harmonic Approximation. *Trends in Physical Chemistry* **2006**, *11*, 87-121.
21
22

23 (2) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent
24 Binding. *Chem. Rev.* **2009**, *109*, 4092-4107.
25
26

27 (3) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of Configurational Entropy
28 from Molecular Simulations Via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127*.
29
30

31 (4) Numata, J.; Wan, M.; Knapp, E. Conformational Entropy of Biomolecules:
32 Beyond the Quasi-Harmonic Approximation. *Genome Inform.* **2007**, *18*, 192-205.
33
34

35 (5) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule
36 Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction
37 Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5*, 3150-3160.
38
39

40 (6) Li, D.-W.; Brüschweiler, R. In Silico Relationship between Configurational
41 Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102*,
42 118108.
43
44

45 (7) Hensen, U.; Lange, O. F.; Grubmüller, H. Estimating Absolute Configurational
46 Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS ONE* **2010**, *5*,
47 1-8.
48
49

50 (8) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of
51 Configurational Entropy from Molecular Simulations by Combining the Mutual-Information
52 Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29*, 1605-1614.
53
54

55 (9) Matsuda, H. Physical Nature of Higher-Order Mutual Information: Intrinsic
56 Correlations and Frustration. *Phys. Rev. E* **2000**, *62*, 3098-3102.
57
58
59
60

- 1
2
3 (10) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of
4 Macromolecules. *Macromolecules* **1981**, *14*, 325-332.
5
6 (11) Karplus, M.; Ichiye, T.; Pettit, B. M. Configurational Entropy of Native Proteins.
7 *Biophys. J.* **1987**, *52*, 1083-1085.
8
9 (12) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules
10 Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617-621.
11
12 (13) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance
13 Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289-6292.
14
15 (14) Hensen, U.; Grubmüller, H.; Lange, O. F. Adaptive Anisotropic Kernels for
16 Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional
17 Configuration Spaces. *Phys. Rev. E* **2009**, *80*, 011913.
18
19 (15) Strajbl, M.; Sham, Y. Y.; Villà, J.; Chu, Z.-T.; Warshel, A. Calculations of
20 Activation Entropies of Chemical Reactions in Solution. *J. Phys. Chem. B* **2005**, *104*, 4578-4584.
21
22 (16) Singh, N.; Warshel, A. A Comprehensive Examination of the Contributions to
23 Binding Entropy of Protein-Ligand Complexes. *Proteins* **2010**, *78*, 1724-1735.
24
25 (17) Chang, C.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the
26 Quasiharmonic Approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017.
27
28 (18) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley, 2006.
29
30 (19) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K.
31 Configurational Entropy in Protein-Peptide Binding: Computational Study of TSG101 Ubiquitin
32 E2 Variant Domain with an Hiv-Derived Ptap Nonapeptide. *J. Mol. Biol.* **2009**, *389*, 315-335.
33
34 (20) Chang, C. A.; Chen, C.; Gilson, M. K. Ligand Configurational Entropy and
35 Protein Binding. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1534-1539.
36
37 (21) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by
38 Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational
39 Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2011**,
40 *10.1021/ct200216n*.
41
42 (22) DeTar, D. F. Calculation of Entropy and Heat Capacity of Organic Compounds in
43 the Gas Phase. Evaluation of a Consistent Method without Adjustable Parameters. Applications
44 to Hydrocarbons. *J. Phys. Chem. A* **2007**, *111*, 4464-4477.
45
46 (23) Suárez, E.; Díaz, N.; Suárez, D. Entropic Control of the Relative Stability of
47 Triple-Helical Collagen Peptide Models. *J. Phys. Chem. B* **2008**, *112*, 15248-15255.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 (24) Gohlke, H.; Case, D. A. Converging Free Energy Estimates: MM-PB(GB)SA
4 Studies on the Protein–Protein Complex Ras–Raf. *J. Comput. Chem.* **2003**, *25*, 238-250.

5
6 (25) Turk, B. E.; Huang, L. L.; E.T., P.; Cantley, L. C. Determination of Protease
7 Cleavage Site Motifs Using Mixture-Based Oriented Peptide Libraries. *Nat. Biotechnol.* **2001**,
8 *19*, 661-667.

9
10 (26) Suárez, E.; Suárez, D. Multibody Local Approximation: Application to
11 Conformational Entropy Calculations on Biomolecules (Submitted). **2011**.

12
13 (27) Overall, C. M.; Kleifeld, O. Validating Matrix Metalloproteinases as Drug Targets
14 and Anti-Targets for Cancer Therapy. *Nature Rev. Cancer* **2006**, *6*, 227-239.

15
16 (28) Smith, M. M.; Shi, L.; Navre, M. Rapid Identification of Highly Active and
17 Selective Substrates for Stromelysin and Matrilysin Using Bacteriophage Peptide Display
18 Libraries. *J. Biol. Chem.* **1995**, *270*, 6440-6449.

19
20 (29) Kolossváry, I.; Guida, W. C. Low-Mode Conformational Search Elucidated:
21 Application to C₃₉H₈₀ and Flexible Docking of 9-Deazaguanine Inhibitors into PNP. *J. Comput.*
22 *Chem.* **1999**, *20*, 1671–1684.

23
24 (30) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke,
25 R. E.; Luo, R.; Crowley, M.; R.C.Walker; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.;
26 Roitberg, A.; Seabra, G.; I.Kolossváry; K.F.Wong; Paesani, F.; Vanicek, J.; X.Wu; Brozell, S. R.;
27 Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D.
28 H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A.; AMBER10 University of California: San
29 Francisco, 2008.

30
31 (31) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.;
32 Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for
33 Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical
34 Calculations. *J. Comput. Chem.* **2003**, *14*, 1999.

35
36 (32) Brown, R. A.; Case, D. A. Second Derivatives in Generalized Born Theory. *J.*
37 *Comput. Chem.* **2006**, *27*, 1662–1675.

38
39 (33) Case, D. A.; Cheatham, T. E.; Darden, T.; Gholke, H.; R., L.; Merz, K. M.;
40 Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The Amber Biomolecular Simulation
41 Programs. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

42
43 (34) Browner, M. F.; Smith, W. W.; Castelhanol, A. L. Matrilysin-Inhibitor
44 Complexes: Common Themes among Metalloproteases. *Biochemistry* **1995**, *34*, 6602-6610.

1
2
3 (35) Diaz, N.; Suarez, D. Molecular Dynamics Simulations of the Active Matrix
4 Metalloproteinase-2: Positioning of the N-Terminal Fragment and Binding of a Small Peptide
5 Substrate. *Proteins* **2008**, *2008*, 50-61.
6
7

8 (36) Díaz, N.; Suarez, D.; Suarez, E. Kinetic and Binding Effects in Peptide Substrate
9 Selectivity of Matrix Metalloproteinase-2: Molecular Dynamics and QM/MM Calculations.
10 *Proteins* **2010**, *78*, 1-11.
11
12

13 (37) Diaz, N.; Suárez, D.; Valdes, H. From the X-Ray Compact Structure to the
14 Elongated Form of the Full-Length Mmp-2 Enzyme in Solution: A Molecular Dynamics Study.
15 *J. Am. Chem. Soc.* **2008**, *130*, 14070-14071.
16
17

18 (38) Case, D. A.; Darden, T. A.; Cheatham, I., T.E. ; Simmerling, C. L.; Wang, J.;
19 Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.;
20 Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.;
21 Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.;
22 Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.;
23 Kovalenko, A.; Kollman, P. A.; AMBER 11 ed.; University of California, Ed. San Francisco,
24 2010.
25
26
27
28
29
30

31 (39) Taylor, C. C. Automatic Bandwidth Selection for Circular Density Estimation.
32 *Comput. Stat. Data Anal.* **2008**, *52*, 3493-3500.
33
34

35 (40) Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation*
36 **2003**, *15*, 1191-1253.
37

38 (41) *Free Energy Calculations*; Chipot, C.; Pohorille, A., Eds.; Springer-Verlag: Berlin
39 Heidelberg, 2007.
40
41

42 (42) Almlöf, M.; Carlsson, J.; Aqvist, J. Improving the Accuracy of the Linear
43 Interaction Energy Method for Solvation Free Energies. *J. Chem. Theory Comput.* **2007**, *3*, 2162-
44 2175.
45
46

47 (43) Sham, Y.-Y.; Chu, Z.-T.; Tao, H.; Warshel, A. Examining Methods for
48 Calculations of Binding Free Energies: LRA, LIE, PDL-D-LRA, and PDL-D/S-LRA Calculations
49 of Ligands Binding to an HIV Protease. *Proteins* **2000**, *39*, 393-407.
50
51

52 (44) Klepeis, J. L.; Lindorff-Larsen, K.; R.O., D.; Shaw, D. E. Long-Timescale
53 Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Op. Struct. Biol.* **2009**,
54 *19*, 120-127.
55
56
57
58
59
60

CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations

SUPPORTING INFORMATION

Ernesto Suárez,* Natalia Díaz, Jefferson Méndez and Dimas Suárez*

Departamento de Química Física y Analítica. Universidad de Oviedo.

C/ Julián Clavería, 8. 33006, Oviedo. Spain.

Phone: +34-985103689

Fax: +34-985103125

E-mail: dimas@uniovi.es

E-mail: ernesto@fluor.quimica.uniovi.es

Table S1. Conformational entropy values (in cal/(K mol)) and computer CPU time^a (in hours; values in italics) for different methods implemented in CENCALC. All the calculations were done on snapshots taken from the 1.0 μ s MD trajectory of the **GNR** peptide and using a cutoff value of 8 Å.

<i>Num of Snapshots /10³</i>	MIE		AMIE		MLA		CC-MLA	
	S_{conform}	CPU time	S_{conform}	CPU time	S_{conform}	CPU time	S_{conform}	CPU time
50	36.48	8.87	36.71	<i>1.18</i>	34.37	<i>0.002</i>	40.35	<i>0.04</i>
100	35.33	<i>20.84</i>	35.38	2.98	36.40	<i>0.005</i>	41.93	<i>0.14</i>
150	36.78	<i>33.56</i>	37.38	<i>5.01</i>	37.84	<i>0.01</i>	43.93	<i>0.32</i>
200	38.27	<i>48.18</i>	38.54	<i>7.50</i>	39.09	<i>0.02</i>	43.99	<i>0.49</i>
250	39.47	<i>62.29</i>	39.56	<i>9.93</i>	39.95	<i>0.03</i>	44.35	<i>0.69</i>
300	39.57	<i>77.03</i>	39.26	<i>12.45</i>	40.79	<i>0.04</i>	45.09	<i>0.94</i>
350	40.19	<i>93.19</i>	39.79	<i>15.30</i>	41.36	<i>0.06</i>	45.35	<i>1.22</i>
400	41.31	<i>111.23</i>	40.98	<i>18.47</i>	41.72	<i>0.08</i>	45.61	<i>1.57</i>
450	41.73	<i>128.40</i>	41.45	<i>21.67</i>	41.94	<i>0.10</i>	45.79	<i>1.94</i>
500	44.04	<i>145.79</i>	44.36	<i>24.80</i>	42.28	<i>0.11</i>	47.00	<i>2.84</i>
550	44.17	<i>165.63</i>	44.58	<i>28.64</i>	42.51	<i>0.15</i>	46.88	<i>3.27</i>
600	41.33	<i>187.06</i>	41.57	<i>32.62</i>	42.68	<i>0.16</i>	47.44	<i>4.14</i>
650	42.66	<i>208.31</i>	42.81	<i>36.89</i>	42.95	<i>0.19</i>	47.76	<i>4.87</i>
700	42.84	<i>228.97</i>	43.16	<i>40.68</i>	43.02	<i>0.20</i>	48.28	<i>6.18</i>
750	42.79	<i>250.35</i>	42.98	<i>44.83</i>	43.22	<i>0.22</i>	48.38	<i>7.15</i>
800	42.82	<i>272.17</i>	43.01	<i>49.00</i>	43.44	<i>0.25</i>	48.82	<i>8.55</i>
850	43.27	<i>294.28</i>	43.44	<i>53.56</i>	43.63	<i>0.31</i>	48.73	<i>9.24</i>
900	43.53	<i>317.87</i>	43.65	<i>58.30</i>	43.88	<i>0.36</i>	48.76	<i>10.23</i>
1000	43.51	<i>340.64</i>	43.54	<i>63.05</i>	44.08	<i>0.42</i>	48.76	<i>10.26</i>

(a) For a Xeon 5450 single core.

Figure S1 Convergence plots for the entropy calculations summarized in Table S1.

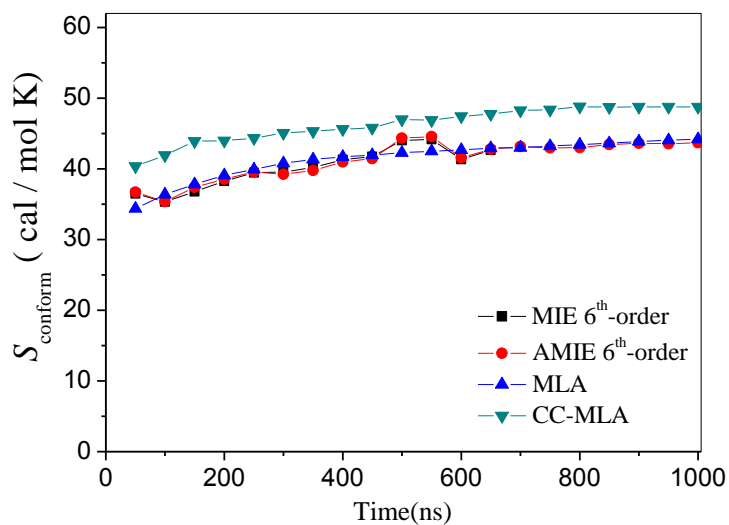


Figure S2. AMIE and CC-MLA entropy plots for the unbound form of the GNR peptide discriminating between total, backbone and side chain contributions.

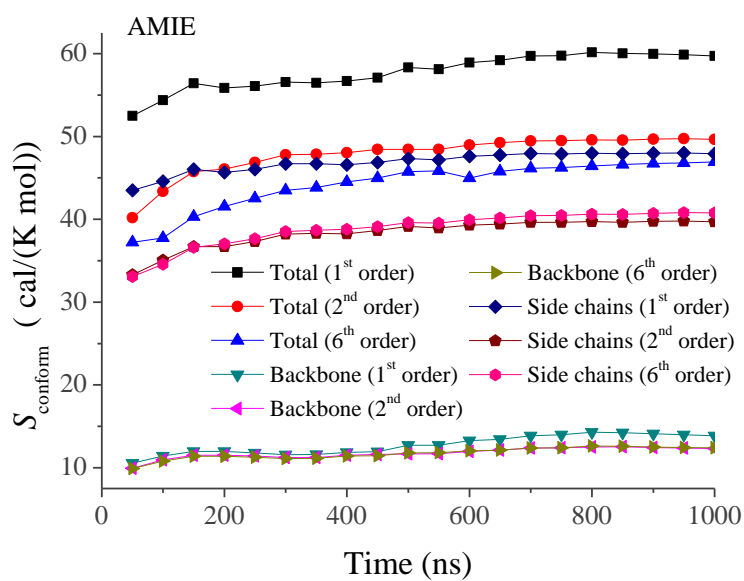


Figure S2.(cont.)

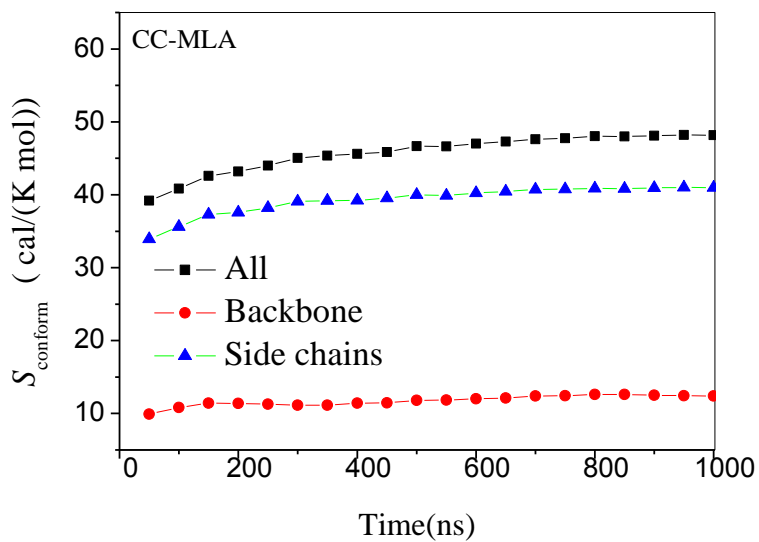


Figure S3 Comparison between the MLA and CC-MLA entropies for the 0.4 μ s trajectory of the MMP7-bound GNR peptide at different cutoff values.

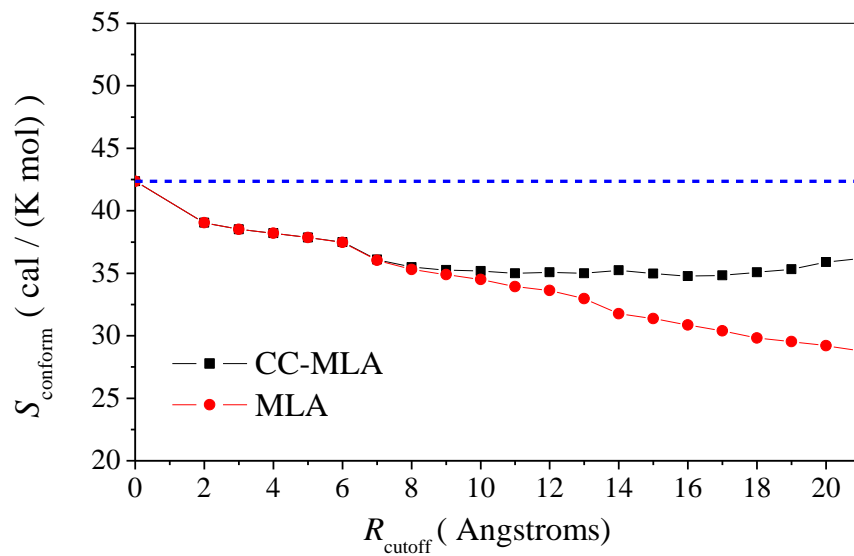
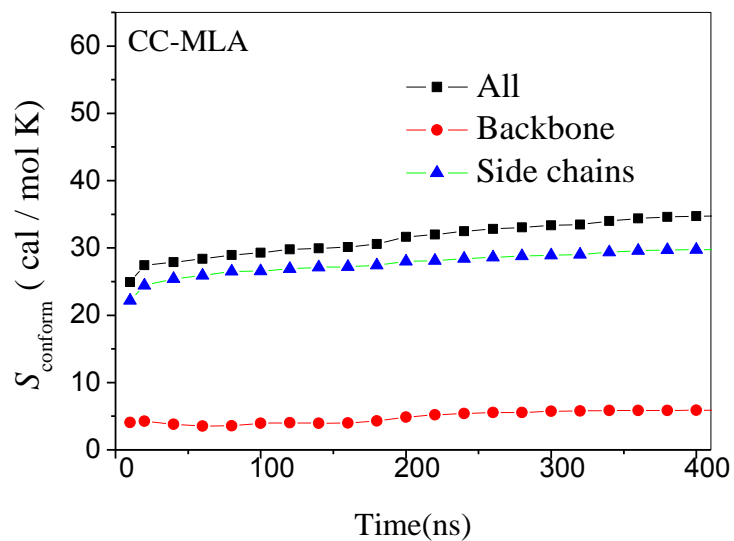
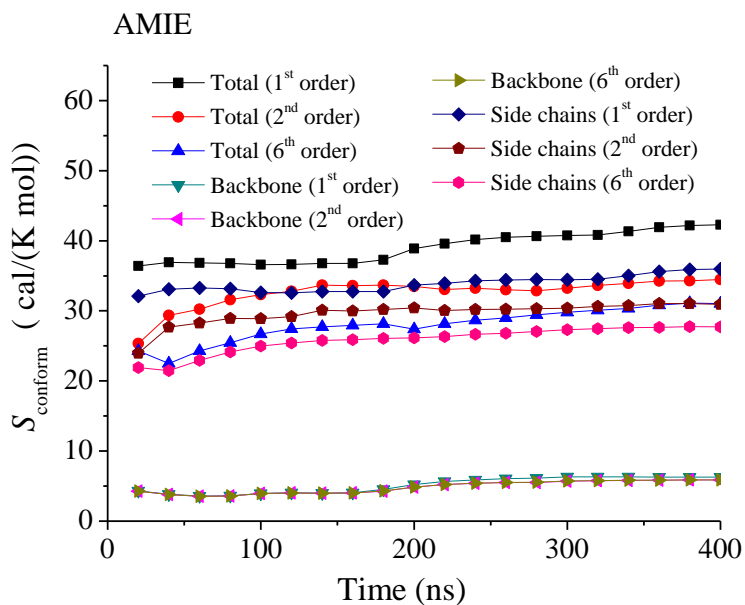


Figure S4. AMIE and CC-MLA entropy plots for the MMP7-form of the **GNR** peptide discriminating between total, backbone and side chain contributions.



2.1.2 Otros Cálculos de Entropía Conformacional en Modelos de Colágeno

Con el fin de complementar los resultados mostrados en los artículos y manuscritos relativos al cálculo de la entropía conformacional, en esta sección se mostrarán resultados adicionales obtenidos utilizando nuestra última propuesta metodológica, denominada CC-MLA (trabajo 2.1.1.5), en varios modelos de colágeno. Para los sistemas ya estudiados POG10 y T3-785, se compararan los nuevos resultados con los ya publicados en nuestro primer trabajo 2.1.1.1. Consideraremos además dos nuevos modelos de triple hélice, THP-1 y fTHP5 descritos en la introducción general de la Tesis, para los que se discutirán brevemente las particularidades de su estructura y dinámica, y se analizarán los cambios en la entropía conformacional asociados a la formación de la triple hélice.

Todas las simulaciones, tanto de las formas en triple hélice como de las hebras que las constituyen aisladas o desnaturalizadas, emplearon el campo de fuerzas AMBER03.⁽¹²¹⁾ Se utilizó disolvente explícito con el modelo TIP3P,⁽¹²²⁾ condiciones periódicas de contorno a 300 K y presión 1.0 atm y correcciones PME para las interacciones de largo alcance.⁽¹²³⁾ Las ecuaciones de Newton se integraron utilizando un paso de 2.0 fs, restringiendo los enlaces que involucran átomos ligeros con el algoritmo SHAKE.⁽¹²⁴⁾

Las estructuras iniciales de las triples hélices desnaturalizadas (hebras aisladas en los modelos POG10, T3-785 y fTHP-5) se generaron con la ayuda del programa LMOD incluido en el paquete AMBER. Esta búsqueda conformacional, basada en el análisis de los modos normales de vibración de baja frecuencia, se realizó en presencia del modelo continuo de disolvente GB de Hawkins-Cramer-Truhlar (GB-HCT).⁽¹²⁵⁾ Para el sistema THP-1, en el que las tres cadenas están unidas por enlaces covalentes, se generó una primera forma desnaturalizada mediante simulaciones dinámicas GB-HCT, empleando inicialmente una temperatura elevada, 500 K, que se disminuyó gradualmente hasta los 300 K. La estructura desordenada del THP-1 así generada se refinó con cálculos LMOD. Los tiempos totales de simulación en disolvente explícito, que varían en cada caso según el tamaño del sistema, los recursos computacionales

disponibles y el grado de convergencia de los cálculos de entropía, serán especificados más adelante para cada modelo.

2.1.2.1 Modelos POG10 y T3-785

En nuestro primer trabajo (2.1.1.1) ya analizamos la estructura y dinámica de los modelos POG10 y T3-781, tanto en su forma de triple hélice como de hebras disociadas. Por tanto, en este apartado nos centraremos únicamente en presentar cómo se aplica el nuevo protocolo de cálculo de entropías conformacionales CC-MLA en ambos casos y comprobar si se mantienen los resultados obtenidos con anterioridad.⁽²⁶⁾ Los tiempos de simulación de estos modelos fueron de 15 ns para los estados de triple hélice y de 15 ns y 250 ns para los estados disociados de los modelos POG10 y T3-785, respectivamente. La razón de este contraste tan grande es una gran diferencia en flexibilidad. La hebra disociada del modelo T3-785 tiene menos anillos de Prolina e Hydroxiprolina, y no está formando una estructura rígida de triple hélice, consecuentemente, es mucho más flexible que el resto de los modelos. Por esta razón tuvimos que hacer, en su momento, un gran esfuerzo computacional para obtener valores razonablemente convergidos de la entropía conformacional del sistema T3-785.

Modelo POG10

Para aplicar el último protocolo propuesto necesitamos saber cuál es el *cutoff* óptimo dada la naturaleza de nuestro sistema y el tamaño de muestra (número de estructuras o *snapshots*). El *cutoff* óptimo R^* será aquel que minimice la entropía CC-MLA ya que será el que nos dará la menor de las cotas superiores que podemos estimar con la información dada (véase el trabajo 2.1.1.5). La siguiente figura muestra la dependencia con el *cutoff* (que simbolizamos con R) de la entropía MLA y CC-MLA para el modelo POG10 en su forma disociada (Figura 2.1A) y de triple hélice (Figura 2.1B).

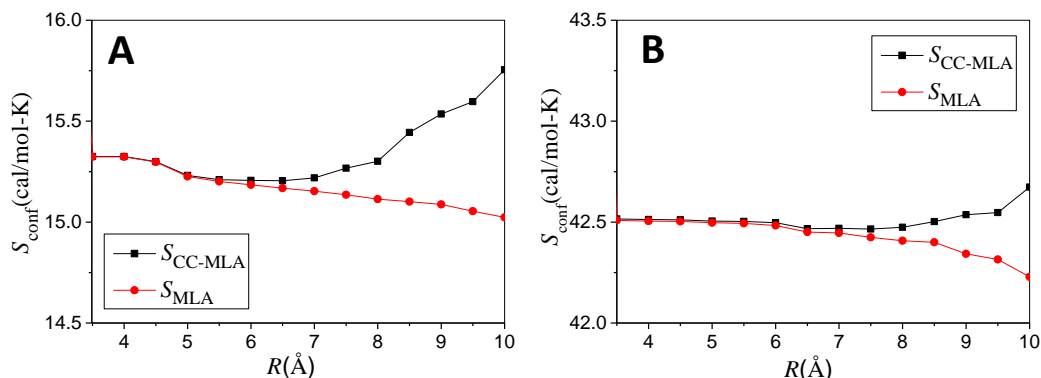


Figura2.1: Entropía conformacional MLA y CC-MLA evaluada a diferentes valores de cutoff (R) en el modelo POG10, tanto en su forma de hebra aislada (A) como de triple hélice (B).

En este caso los mínimos CC-MLA coinciden alrededor de $R^* = 6.5 \text{ \AA}$, que sería el valor óptimo según nuestro protocolo. Conocido éste, podemos entonces visualizar la evolución temporal de la entropía estimada para ambos modelos (Figura 2.2). En la figura se muestra además las contribuciones a la entropía de la cadena principal, de las cadenas laterales y la suma de entropías marginales (SEM) ya que su convergencia es punto de partida de la aproximación CC-MLA.

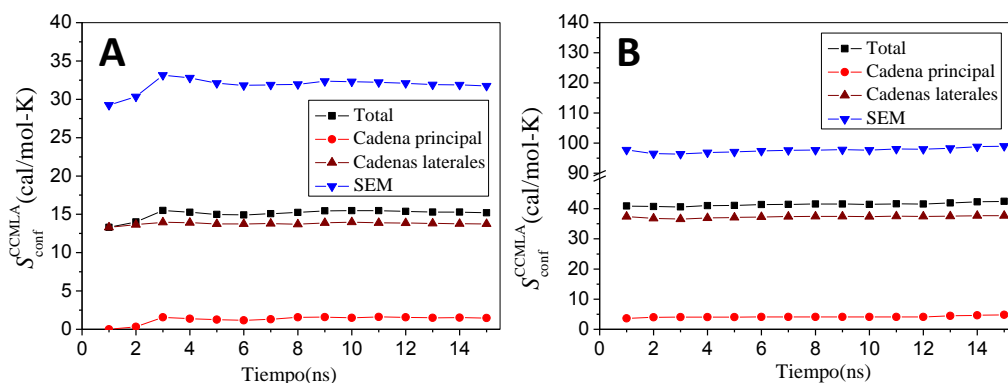


Figura2.2: Evolución de la entropía conformacional en el tiempo evaluada a $R^* = 6.5 \text{ \AA}$ en el modelo POG10, tanto en su forma de hebra aislada (A) como de triple hélice (B). En ambos casos se presentan las contribuciones de la cadena principal y de las cadenas laterales a la entropía total. También se muestra la evolución de la suma de entropías marginales (SEM).

Si comparamos los resultados obtenidos por el método CC-MLA con los proporcionados por el método el propuesto originalmente(26) encontramos ligeras diferencias. Recordemos que entonces los cálculos sólo se hacían sobre la cadena principal, que es por otro lado, todo cuanto podíamos hacer con el método utilizado (MIE) y la extensión de las simulaciones. El valor de $-T\Delta S_{conform}$ en la formación de la triple hélice fue en nuestro primer trabajo de 0.4 kcal por mol de péptido (véase la Tabla

3 en 2.1.1.1), mientras que si repetimos el cálculo utilizando CC-MLA sobre las cadenas principales obtenemos 0.0 kcal por mol de péptido. Ambos valores nos llevan a la misma conclusión: la gran presencia de anillos de prolina e hidroxiprolina prácticamente anula el impacto de esta magnitud en la energía libre de formación de la triple hélice. Incluyendo el efecto de las cadenas laterales, gracias a las mejoras introducidas por el método CC-MLA, obtenemos un valor de $-T\Delta S_{conf}$ igual a 0.3 kcal/mol, que apenas se diferencia del cálculo hecho exclusivamente sobre la cadena principal en este caso particular.

Modelo T3-785

En este modelo, los valores óptimos de cutoff son de 8.5 Å para la hebra disociada y de 9.0 Å para la triple hélice. La Figura 2.3 muestra las gráficas de la dependencia con el *cutoff* de la entropía conformacional MLA y CC-MLA para los estados de hebra disociada y de triple hélice.

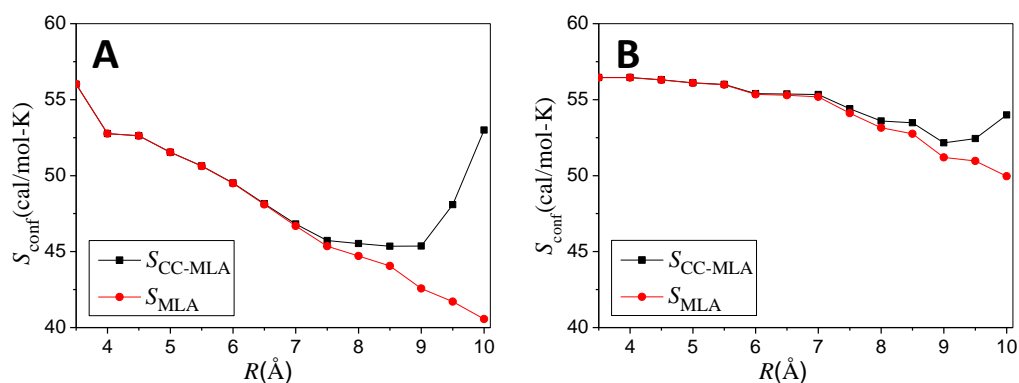


Figura 2.3: Entropía conformacional MLA y CC-MLA evaluada a diferentes valores de *cutoff* (R) en el modelo T3-785, tanto en su forma de hebra aislada (A) como de triple hélice (B).

Con los valores obtenidos podemos representar entonces la evolución temporal de los valores estimados CC-MLA de entropía (ver Figura 2.4).

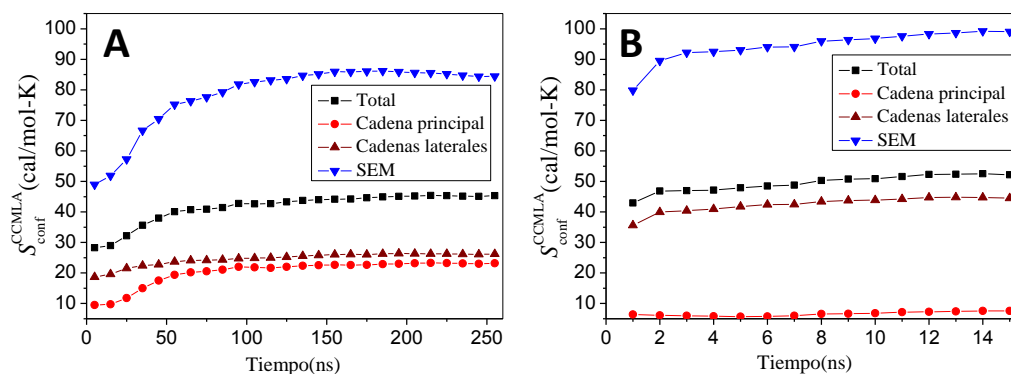


Figura 2.4: Evolución de la entropía conformacional con el tiempo evaluada para el valor de R óptimo en el modelo T3-785, tanto en su forma de hebra aislada ($R^* = 8.5 \text{ \AA}$, A) como de triple hélice ($R^* = 9.0 \text{ \AA}$, B). En ambos casos se presentan las contribuciones de la cadena principal y de las cadenas laterales a la entropía total. También se muestra la evolución de la suma de entropías marginales (SEM).

Si comparamos estos últimos resultados en el caso de la cadena principal ($-T\Delta S_{conform} = 6.2$ kcal/mol de péptido) con los obtenidos en nuestro primer trabajo ($-T\Delta S_{conform} = 4.3$ kcal/mol de péptido) observamos una diferencia de apenas 1.6 kcal/mol. A este respecto, es de esperar que los nuevos resultados obtenidos con el método CC-MLA proporcionen una mejor estimación del cambio en la entropía conformacional asociada a la formación de la triple hélice T3-785, debido a que elimina la incertidumbre asociada al truncamiento en el orden y filtra la falsa correlación (véase el trabajo 2.1.1.5). Este cambio de entropía pasa a ser 8.4 kcal/mol si añadimos el efecto de las cadenas laterales, lo que acentúa aún más la diferencia de estabilidad debida a la entropía conformacional. Sin embargo, tal y como se aprecia en la Figura 2.4B, la SEM está menos convergida para el estado de triple hélice, por lo que este último valor podría estar sobreestimando la diferencia de entropía. Ello sugeriría la necesidad de extender en el tiempo la simulación de dinámica molecular de la triple hélice T3-785 inicialmente realizada, o realizar simulaciones adicionales con distintas condiciones iniciales que nos permitiesen acumular estructuras minimizando el problema de la interacción entre las imágenes presente en este tipo de sistemas alargados.

2.1.2.2 Modelo THP-1

En este modelo (sección 1.3.3), los tiempos de simulación fueron aproximadamente de unos 40 ns para la triple hélice y de alrededor de 400 ns para el estado no colagénico.

Hay que señalar que en este caso particular del THP-1, la presencia de enlaces covalentes entre las tres hebras determina que la forma desnaturalizada del sistema tenga el mismo tamaño (mismo número de residuos) que la forma de triple hélice. Esto supone tener un espacio conformacional mucho mayor que en modelos como el POG10 o el T3-785, donde los estados disociados los podemos representar con una sola hebra. Por esa razón decidimos extender mucho más la simulación de la forma desnaturalizada del sistema THP-1 en comparación con el modelo de triple hélice.

En la dinámica molecular del estado de triple hélice se verifica el mismo tipo de contactos entre cadenas previamente observado para los sistemas POG10 y T3-785: puentes de hidrógeno directos entre cadenas, así como puentes de hidrógeno mediados por moléculas de agua. La Figura 2.5 muestra un fragmento del modelo THP-1 donde representamos los puentes de hidrógeno directos con líneas discontinuas y se resaltan algunas moléculas de agua que estarían asistiendo interacciones entre distintas partes de la proteína.

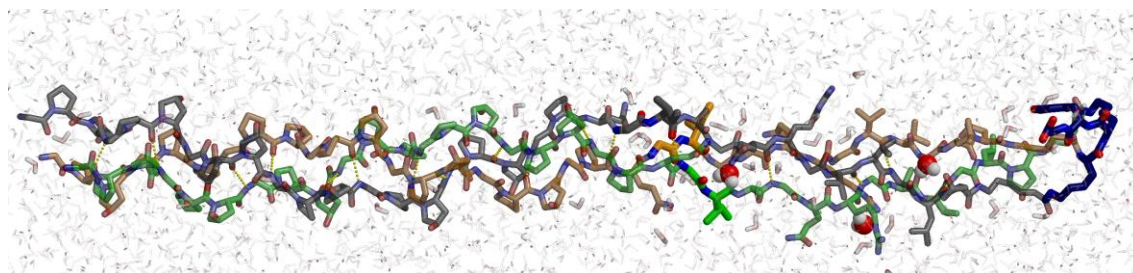


Figura 2.5: Estructura representativa del modelo THP-1. Se muestra los puentes de hidrógeno directos entre cadenas así como aquellas moléculas de agua (en formato CPK) que estarían asistiendo interacciones por puentes de hidrógeno.

Estos contactos se pueden agrupar formando patrones muy característicos que son estables a lo largo de toda la simulación. La siguiente Tabla muestra como ejemplo los puentes de hidrógeno de la zona central y no prototípica del modelo, caracterizándolos por su abundancia y distancia promedio (criterios geométricos para considerar la existencia de puente de hidrógeno: distancia entre los átomos pesados $< 3.5 \text{ \AA}$ y el ángulo H-donor-aceptor $< 60^\circ$).

Puentes de Hidrógeno	THP-1	
	%	distancia(Å)
Gly ₁₉ -NH...OC-Pro ₅₁	94.72	3.10±0.17
Gly ₂₂ -NH...OC-Pro ₅₄	97.12	3.08±0.16
Gly ₂₅ -NH...OC-Ile ₅₇	99.58	2.92±0.14
Gly ₂₈ -NH...OC-Gln ₆₀	99.61	2.95±0.14
Gly ₃₁ -NH...OC-Val ₆₃	99.50	2.98±0.15
Gly ₅₃ -NH...OC-Pro ₈₉	94.09	3.11±0.17
Gly ₅₆ -NH...OC-Pro ₉₂	97.14	3.09±0.16
Gly ₅₉ -NH...OC-Ile ₉₅	99.29	2.92±0.14
Gly ₆₂ -NH...OC-Gln ₉₈	99.90	2.93±0.13
Gly ₆₅ -NH...OC-Val ₁₀₁	98.96	3.01±0.16
Gly ₈₈ -NH...OC-Pro ₁₇	97.04	3.06±0.17
Gly ₉₁ -NH...OC-Pro ₂₀	97.48	3.02±0.17
Gly ₉₄ -NH...OC-Ile ₂₃	97.92	3.05±0.16
Gly ₉₇ -NH...OC-Gln ₂₆	98.59	2.98±0.16
Gly ₁₀₀ -NH...OC-Val ₂₉	98.92	3.02±0.16

Tabla2.1: Puentes de hidrógeno directos entre cadenas. El grupo N-H de las glicinas interacciona con el grupo C=O de la posición X más cercana de la hebra vecina. La distancia está medida entre los átomos pesados involucrados.

También se observan los patrones de puentes de hidrógeno mediados por moléculas de agua que suelen verse en experimentos de Rayos-X.(81, 91) Los carbonilos de glicinas interactúan a través una molécula de agua con un residuo en posición X de una cadena adyacente siempre que esta posición no esté ocupada por un imido-ácido. La Tabla 2.2 agrupa los puentes de hidrógeno con estas características que superan el 50% de abundancia a lo largo de la dinámica.

THP-1	Abundancia
Puentes de H mediados por H ₂ O	%
Gly ₂₂ -CO...HO...HN-Ile ₅₇	82.22
Gly ₂₅ -CO...HO...HN-Gln ₆₀	65.22
Gly ₂₈ -CO...HO...HN-Val ₆₃	89.68
Gly ₃₁ -CO...HO...HN-Leu ₆₆	81.11
Gly ₉₁ -CO...HO...HN-Ile ₂₃	83.65
Gly ₉₄ -CO...HO...HN-Gln ₂₆	79.79
Gly ₉₇ -CO...HO...HN-Val ₂₉	81.64

Tabla2.2: Puentes de hidrógeno mediados por agua con abundancia superior al 50%. Nótese que se trata de puentes "lineales", sólo participa uno de los protones del agua en el puente.

La helicidad que es la propiedad estructural más característica de la triple hélice, la cuantificamos mediante el ángulo de giro definido en 2.1.1.1. Como vimos entonces, el grado de helicidad varía considerablemente a lo largo una secuencia heterogénea. En la

Figura 2.6 vemos la variación de esta magnitud en la triple hélice del modelo THP-1. Como era de esperar, la primera mitad de la triple hélice con secuencia prototípica tiene un mayor ángulo de giro, perdiéndose en parte la helicidad en la segunda mitad donde el contenido de imido-ácidos es menor.

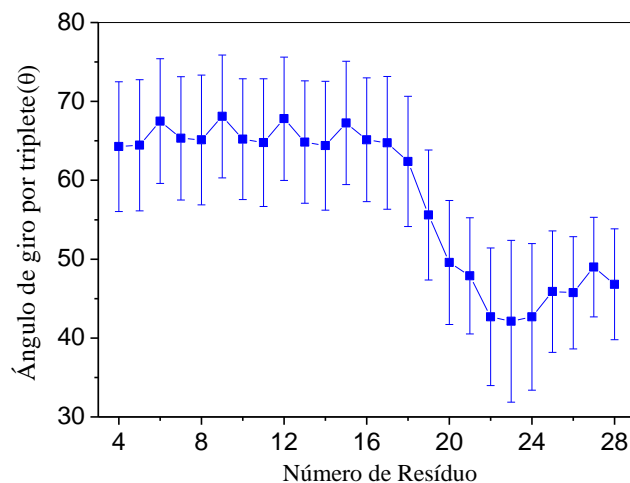


Figura 2.6: Ángulo de giro por triplete promediado sobre las tres cadenas. Las barras verticales muestran la desviación estándar de los resultados.

No se observa una evolución global de estos valores de helicidad, simplemente fluctúan con una desviación típica de $\sim 8^\circ$. La estructura de triple hélice penaliza los cambios conformacionales de las cadenas principales. Aun así, el RMSF (*Root Mean Square Fluctuation*) del *backbone* nos da un valor medio de $2.0 \pm 0.6 \text{ \AA}$ que, al igual que en los modelos POG10 y T3-785, se debe en gran medida a movimientos de plegamiento (*bending*) globales. En este caso también se cumple que más de la mitad de la varianza de las desviaciones con respecto a las estructuras promedio se deben al *bending* global. De hecho, los dos primeros componentes principales obtenidos de la matriz de covarianzas medida sobre los átomos de las cadenas principales son *bending* globales, y juntos explican el 64% de la varianza. Nuestro modelo computacional nos predice que el THP-1 adopta en efecto una estructura en triple hélice, compartiendo propiedades y variabilidad estructural como la observada experimental y teóricamente en otros modelos de colágeno más simple como el POG10 y el T3-785.

A diferencia de las formas colagénicas, los estados desestructurados (*unfolded*) son sistemas más flexibles que no tienen que cumplir con las restricciones que impone una estructura de triple hélice. Este estado pasa por diferentes conformaciones, más o menos

extendidas, que podemos monitorizar a través de la evolución del radio de giro R_{gyr} representado en la Figura 2.7. En la misma figura se muestra un análisis de *clustering* hecho con ayuda de las herramientas MMTSB (126) utilizando los diedros de la cadena principal. Hay que tener en cuenta que, en el caso del estado de triple hélice, el mismo análisis (con los mismos criterios) nos daría un solo representante de *cluster*. Es decir, a pesar de la presencia de imido-ácidos en la secuencia, el estado no colagénico muestra una cierta flexibilidad, hecho que sugiere un papel importante de la entropía conformacional en la formación de la triple hélice.

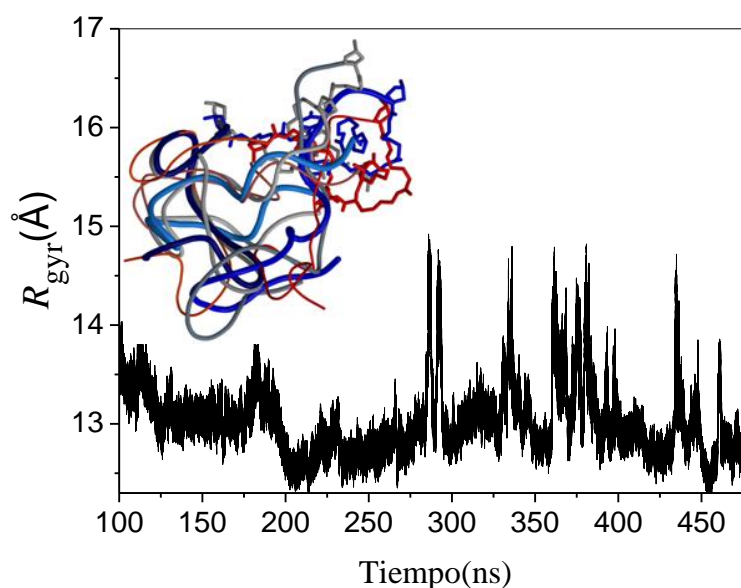


Figura 2.7: Evolución del radio de giro durante la dinámica molecular de la forma no colagénica del THP-1. Se muestra, además, la superposición de un conjunto de estructuras representativas (representantes de cluster) para este modelo. Para el clustering, se utilizan los diedros del backbone y para el dibujo de cada representante se emplea un grosor proporcional a su población.

Calcularemos ahora la entropía conformacional de ambas formas del THP-1, triple hélice y desestructurada, así como el cambio de la misma al formarse la triple hélice. En este caso, el número de diedros que contribuyen a la entropía conformacional son 304 y 181 para el estado no colagénico y el de triple hélice respectivamente. Según muestra la Figura 2.8, los valores de R óptimos son de 8.0 Å tanto para la triple hélice como para el estado no colagénico.

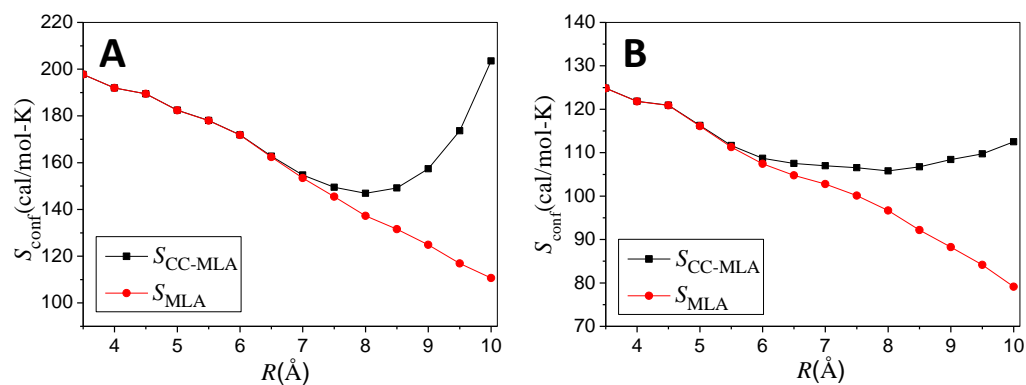


Figura 2.8: Entropía conformacional MLA y CC-MLA evaluada a diferentes valores de cutoff (R) en el modelo THP-1, tanto en su forma no colagénica (A) como de triple hélice (B).

Con los valores obtenidos mostramos en la siguiente figura la evolución temporal de los valores estimados CC-MLA de entropía.

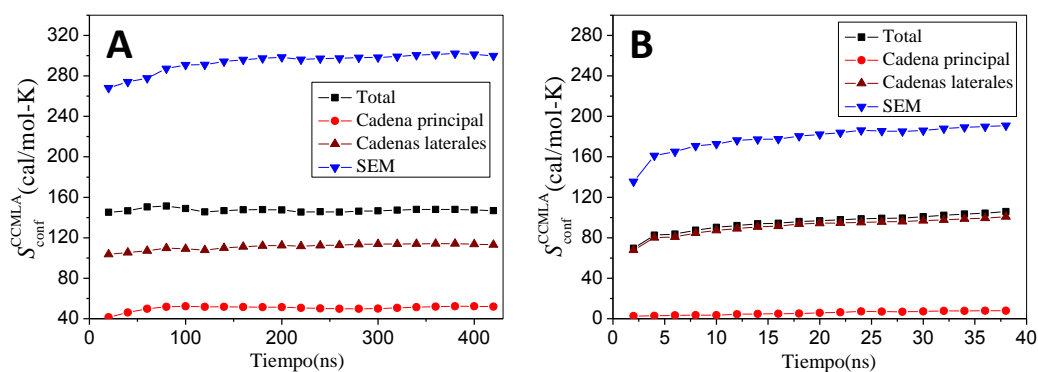


Figura 2.9: Evolución de la entropía conformacional con el tiempo evaluada para un cutoff óptimo $R=8.0$ Å del modelo THP-1, tanto en su forma no colagénica (A) como de triple hélice (B). En ambos casos se presentan las contribuciones de la cadena principal y de las cadenas laterales a la entropía total. También se muestra la evolución de la suma de entropías marginales (SEM).

Lamentablemente en este caso tampoco hay una buena convergencia ni de la SEM ni de la entropía total para el estado de triple hélice. Se trata de otro claro ejemplo donde se necesita mucho más muestreo del espacio configuracional. En cualquier caso, el valor observado de $-T\Delta S_{conform}$ es de 13.2 kcal/mol si utilizamos sólo las torsiones de la cadena principal y de 12.3 kcal/mol si tenemos en cuenta también el efecto de las cadenas laterales.

2.1.2.3 Modelo fTHP-5

En el trabajo 2.1.1.5, donde presentamos el método CC-MLA, se analizó en parte la simulación de una de las hebras aisladas del modelo fTHP-5. En esta sección describiremos brevemente la estructura y el comportamiento dinámico de este modelo, tanto en su forma colagénica como disociada. Finalmente calcularemos la variación de entropía conformacional que tiene lugar asociada a la formación de la triple hélice siguiendo el mismo protocolo que en los modelos anteriores.

La Figura 2.10 muestra un fragmento de la zona central del fTHP-5 donde representamos los puentes de hidrógeno directos entre cadenas y algunas de las moléculas de disolvente que participan en puentes de hidrógeno adicionales mediados por agua. Estos son los contactos habituales de la estructura en triple hélice tipo colágeno y han sido caracterizados en todos los modelos.

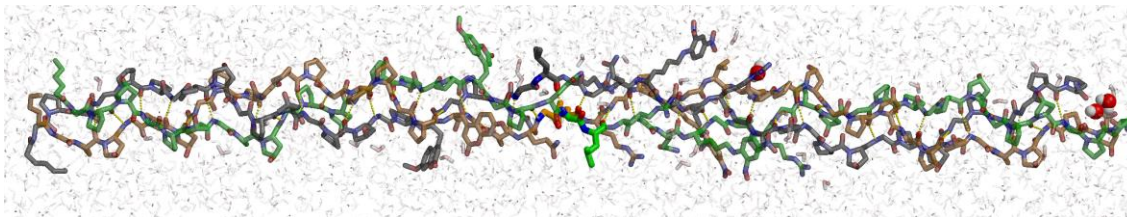


Figura 2.10: Estructura representativa del modelo fTHP-5. Se muestra los puentes de hidrógeno directos entre cadenas así como moléculas de agua que se estarían asistiendo otras interacciones por puentes de hidrógeno entre cadenas.

En la Tabla 2.3 cuantificamos algunos de estos puentes de hidrógeno típicos que involucran a las Glicinas de la zona central del modelo, caracterizándolos por su abundancia y distancias promedio.

También podemos caracterizar la abundancia de los puentes de hidrógeno entre distintas cadenas mediados por moléculas de agua, que se recogen en la Tabla 2.4 para aquellos que muestren una abundancia superior al 50%.

Puentes de Hidrógeno	fTHP-5	
	%	distancia(Å)
Gly ₁₇ -NH...OC- Pro ₆₂	87.13	3.13±0.17
Gly ₂₀ -NH...OC- Pro ₆₅	87.42	3.06±0.17
Gly ₂₃ -NH...OC- Pro ₆₈	92.63	3.12±0.17
Gly ₂₆ -NH...OC- Leu ₇₁	99.11	2.98±0.15
Gly ₂₉ -NH...OC- Gln ₇₄	99.96	2.90±0.12
Gly ₃₂ -NH...OC- Val ₇₇	97.01	3.05±0.16
Gly ₆₄ -NH...OC- Pro ₁₀₉	89.62	3.10±0.17
Gly ₆₇ -NH...OC- Pro ₁₁₂	81.55	3.13±0.18
Gly ₇₀ -NH...OC- Pro ₁₁₅	91.52	3.12±0.17
Gly ₇₃ -NH...OC- Leu ₁₁₈	99.43	2.96±0.15
Gly ₇₆ -NH...OC- Gln ₁₂₁	99.79	2.94±0.14
Gly ₇₉ -NH...OC- Val ₁₂₄	98.38	3.04±0.16
Gly ₁₀₈ -NH...OC- Pro ₁₅	86.72	3.12±0.18
Gly ₁₁₁ -NH...OC- Pro ₁₈	87.02	3.11±0.17
Gly ₁₁₄ -NH...OC- Pro ₂₁	88.33	3.07±0.17
Gly ₁₁₇ -NH...OC- Leu ₂₄	97.91	3.02±0.16
Gly ₁₂₀ -NH...OC- Gln ₂₇	99.77	2.94±0.14
Gly ₁₂₃ -NH...OC- Val ₃₀	97.74	3.09±0.16

Tabla2.3: Puentes de hidrógeno directos entre cadenas en la zona no prototípica del modelo fTHP-5. El grupo N-H de las glicinas interacciona con el grupo C=O de la posición X más cercana de la hebra vecina. La distancia está medida entre los átomos pesados involucrados.

Puentes de H mediados por H ₂ O	Abundancia
	%
Gly ₂₃ -CO...HO...HN-Ile ₇₁	66.15
Gly ₂₆ -CO...HO...HN-Gln ₇₄	77.67
Gly ₂₉ -CO...HO...HN-Val ₇₇	91.59
Gly ₁₁₇ -CO...HO...HN-Gln ₂₇	71.14
Gly ₁₂₀ -CO...HO...HN-Val ₃₀	88.43

Tabla2.4: Puentes de hidrógeno mediados por agua con abundancia superior al 50%. Nótese que se trata de puentes "lineales", sólo participa uno de los hidrógenos del agua en el puente.

La helicidad la cuantificamos nuevamente mediante el ángulo de giro definido en 2.1.1.1. En la siguiente figura observamos la variación de esta magnitud a lo largo de la triple hélice del modelo fTHP-5. Se observa claramente la misma tendencia que en los modelos anteriores, de modo que las zonas ricas en imido-ácidos, que en este caso se encuentran en los extremos, muestran una mayor helicidad que la zona central donde el contenido de Prolinas e Hydroxiprolinas es menor.

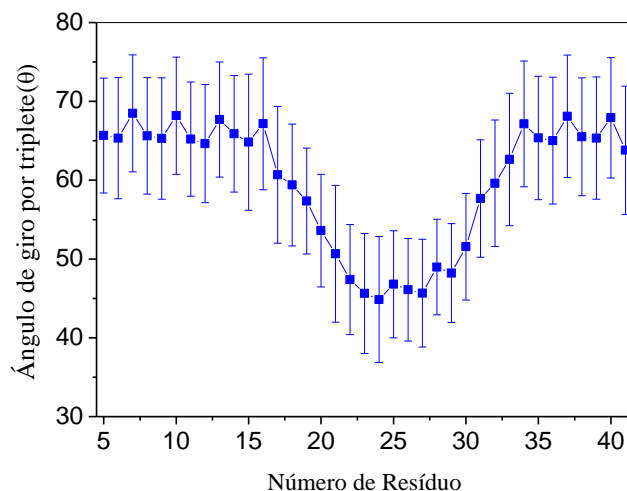


Figura 2.11: Ángulo de giro por triplete promediado sobre las tres cadenas del modelo *fTHP-5*. Las barras verticales muestran la desviación estándar de los resultados.

Tampoco en este caso se observa una evolución global de estos valores de helicidad, simplemente fluctúan a lo largo de la dinámica con una desviación típica de $\sim 7.5^\circ$. Para este modelo, el RMSF del *backbone* nos da un valor medio de $3.8 \pm 0.4 \text{ \AA}$ asociado en gran medida a movimientos de doblaje (*bending*) globales. Nuevamente más de la mitad de la varianza de las desviaciones con respecto a las estructuras promedio se deben al *bending* global. Para este caso, los dos primeros componentes principales obtenidos utilizando los átomos de las cadenas principales son *bending* globales, y juntos explican el 68% de la varianza.

En contraste con el estado de triple hélice esencialmente rígido que acabamos de caracterizar, el estado disociado del *fTHP-5* muestra una gran flexibilidad. La siguiente figura muestra la evolución temporal del radio de giro y la representación gráfica del análisis de *clustering* de la hebra disociada. Como veremos a continuación, esta diferencia de flexibilidad penaliza la formación de la triple hélice en unas cuantas kcal/mol.

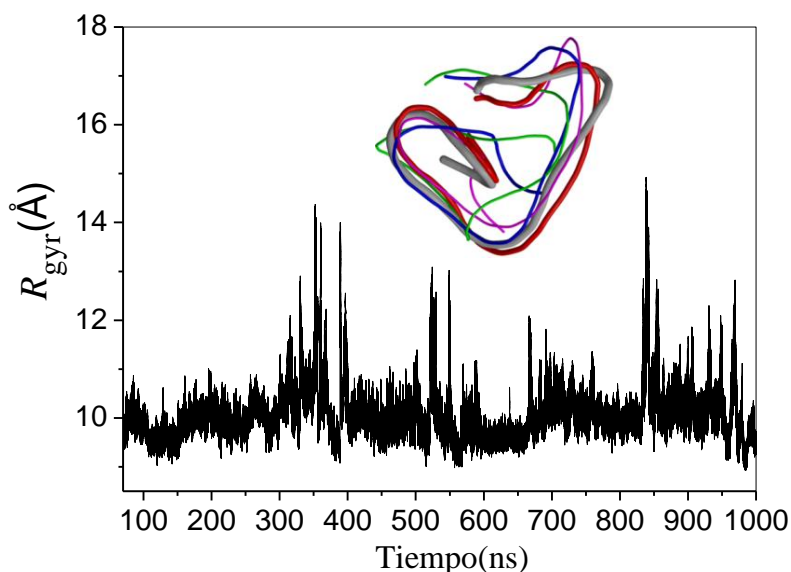


Figura2.12: Evolución del radio de giro durante la dinámica molecular de la forma no colagénica del *fTHP-5* (hebra aislada). Se muestra además la superposición de un conjunto de estructuras representativas (representantes de clusters) para este modelo. Se utiliza los diedros del backbone para el clustering y en el dibujo de cada representante se emplea un grosor proporcional a su población.

Calculamos seguidamente el cambio en la entropía conformacional asociada a la formación del trímero en su configuración de triple hélice, que pasa de un estado flexible con 3×156 diedros que contribuyen a la entropía conformacional, a otro de triple hélice con 236. En la Figura 2.13 observamos como varían los valores de entropía CC-MLA y MLA a diferentes valores de *cutoff* (R).

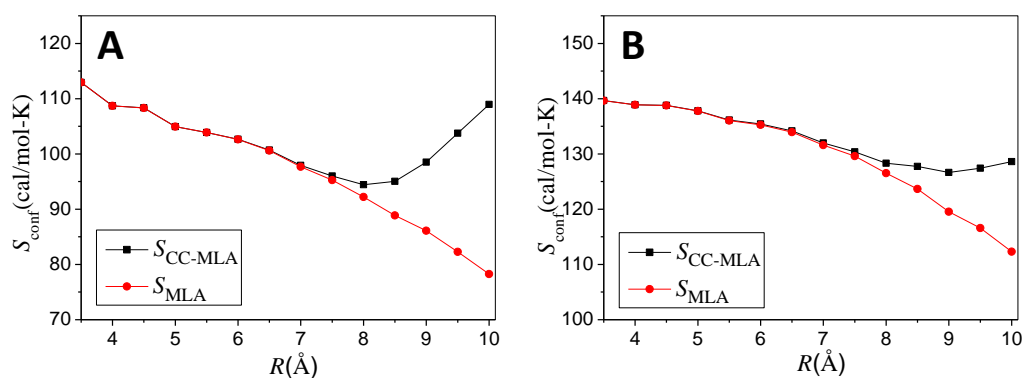


Figura2.13: Entropía conformacional MLA y CC-MLA evaluada a diferentes valores de *cutoff* (R) en el modelo *fTHP-5*, tanto en su forma de hebra aislada (A) como de triple hélice (B).

Los valores de R óptimos (R^*) son en este caso 8.0 \AA para la hebra aislada y 9.0 \AA para la estructura de triple hélice. Con estos resultados mostramos, a continuación, la evolución temporal de los valores estimados de entropía CC-MLA.

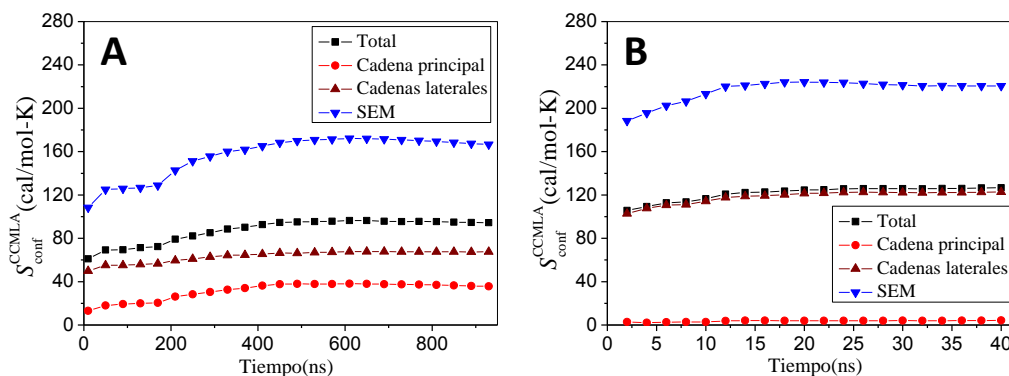


Figura 2.14: Evolución de la entropía conformacional evaluada para el valor de R óptimo en el modelo *fTHP-5*, tanto en su forma de hebra aislada con $R=8.0\text{Å}$ (A) como de triple hélice ($R=9.0\text{Å}$) (B). En ambos casos se muestran las contribuciones de la cadena principal y de las cadenas laterales a la entropía total. También se recoge la evolución temporal de la suma de entropías marginales (SEM).

En este último ejemplo, la contribución de la entropía conformacional a la variación de energía libre $-T\Delta S_{conf}$ es igual a 10.0 kcal/mol teniendo en cuenta solamente la cadena principal, y 15.7 kcal/mol si incluimos la contribución de las cadenas laterales.

2.1.2.4 Observaciones Generales

Con todos los ejemplos anteriores de cálculos de entropía conformacional sobre modelos simulados en la presente Tesis, hemos visto como el valor de R óptimo, que simbolizamos con R^* , depende no sólo del número de puntos utilizados para su simulación, sino que también tiene una fuerte dependencia con la topología del sistema. Las estructuras de triple hélice tienden a mostrar mayores valores de R^* debido a que estas estructuras tienen una menor “densidad de diedros”. Es decir, para un mismo R , los conjuntos o listas \mathcal{L}_i y $\mathcal{L}_i - \{A_i\}$ definidas en 2.1.1.5 son lógicamente menores en estructuras menos compactas como las triples hélices.

La Tabla 2.5 resume los resultados mostrados en los tres apartados anteriores. Como era de esperar, en estos procesos de asociación, la contribución conformacional al cambio total de entropía se debe en su mayoría al cambio en la flexibilidad de la cadena principal. Aun así, si se toma sólo la contribución de la cadena principal podemos cometer errores importantes como se aprecia claramente en el caso del modelo *fTHP-5*.

Modelo	$-T\Delta S_{conf}$ (kcal/mol)			T_m (°C)
	Cadena principal	Cadenas laterales	Total	
POG10	0.0	0.4	0.3	60
T3-785	6.2	-0.3	8.4	25
THP-1	13.2	3.7	12.3	43
fTHP-5	10.3	8.0	15.7	41

Tabla2.5: Contribución al cambio de energía libre para la formación de la triple hélice asociada a la variación de la entropía conformacional. Se muestran las contribuciones debidas a la cadena principal y a las cadenas laterales calculadas por separado. También se recoge la temperatura media de transición para cada modelo de triple hélice.

Con los pocos datos que tenemos, no podemos asegurar una dependencia clara de la estabilidad de la triple hélice (medida semi-cuantitativamente a través de T_m), con el cambio en la entropía conformacional asociada a su formación. En cualquier caso, sí sabemos ahora que esta componente entrópica contribuye significativamente a la estabilidad de estos sistemas. Evidentemente, deben tenerse en cuenta todas las contribuciones a la energía libre de unión para poder comprender en detalle la estabilidad relativa de las distintas triples hélices. Para los modelos POG10 y T3-785, los cálculos aproximados MM-PBSA reportados en el trabajo 2.2.1 sí que han ofrecido una descripción global compensada. Sin embargo, se ha visto en nuestro laboratorio que la aplicación de este protocolo u otros parecidos a los sistemas THP-1 y fTHP-5 es insatisfactoria debido a varias razones (problemas con la componente no polar de la energía de solvatación, posible influencia del disolvente explícito, etc.) que, en cualquier caso, quedan fuera del ámbito de esta Tesis.

En general somos capaces de reproducir estructuras realistas al menos para los modelos de triple hélice POG10 y T3-785, que son los modelos de los que tenemos alguna información experimental directa. En el caso de las hebras aisladas, sin embargo, las estructuras obtenidas hay que verlas con cierta cautela, ya que es difícil saber si hemos simulado el tiempo suficiente para muestrear al mínimo global de energía libre. Aunque la evolución del radio de giro, los análisis de *clustering* conformacional y las propias curvas de entropía sugieren que las simulaciones aquí presentadas han efectuado un muestreo conformacional amplio (con excepción quizás de la forma libre del THP-1), la necesidad de contar con simulaciones extensas para obtener valores fiables de entropías conformacionales es claramente un punto crítico en todos estos cálculos. Probablemente, este calificativo de “extensas”, dependerá mucho de los recursos computacionales disponibles en el momento de realizar las simulaciones. Pero también

se necesitará más experimentación computacional para perfilar un protocolo que permita diagnosticar mejor la convergencia en la predicción de entropías conformacionales.

A pesar de las propuestas metodológicas presentadas, que van en el sentido de mejorar la estimación de la estabilidad de las THPs en fase acuosa, aún no estamos en condiciones de predecir la estabilidad global o relativa de estos sistemas complejos empleando protocolos aproximados como el método MM-PBSA tal y como hemos señalado anteriormente. Con independencia del debate acerca de la bondad del campo de fuerzas, y de cómo los errores del mismo escalan con el tamaño del sistema,(127) creemos que buena parte de las limitaciones de los métodos aproximados son probablemente debidas al uso de modelos de disolvente continuo para describir la “macromolécula” más compleja configuracionalmente: el disolvente acuoso. Calcular con precisión química (~1 kcal/mol) la energía libre de solvatación en grandes sistemas es un tema sin resolver dentro de la Química Teórica. Para el cálculo de la energía de cavitación, por ejemplo, siguen utilizándose modelos empíricos extraordinariamente simples, que consecuentemente no son ni mucho menos generales.(117-118)

2.2 Cálculos de Energía de Biomoléculas a partir de Fragmentos

En el primer trabajo del segundo tema “*Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide*”, se realiza en primer lugar una revisión de los principales métodos recogidos en la literatura para el cálculo de la energía total de un sistema a partir de la combinación de las energías de sus fragmentos. Como primer resultado significativo, se muestra que el formalismo de MBE (que es totalmente análogo al MIE en el caso de las entropías) establece un marco general para las distintas aproximaciones. Se desarrolla además una nueva propuesta, denominada TFEM (*Thermochemical Fragment Energy Method*) que se basa formalmente en un proceso de fragmentación mediante reacciones químicas. Gracias a ello puede combinarse fácilmente con cualquier modelo continuo de disolvente, al igual que con métodos híbridos QM/MM (*Quantum Mechanical /Molecular Mechanics*).

El esquema desarrollado³ se utiliza entonces para calcular la energía absoluta y la interacción entre cadenas del modelo de triple hélice [Ace-(Pro-Hyp-Gly)₄-Nme]₃, que gracias a su reducido tamaño, puede ser tratado íntegramente mediante determinados métodos QM. Esto nos permitirá realizar una comparación adecuada de las energías obtenidas a partir de nuestra aproximación basada en fragmentos TFEM. Los cálculos QM se realizan sobre un conjunto de 25 estructuras que construimos mediante modelado molecular a partir de coordenadas (*snapshots*) extraídas de la dinámica del modelo POG10. El nivel de cálculo seleccionado combina el funcional tipo GGA no empírico de Perdew-Burke-Ernzerhoff (PBE) con la base SVP (*Split Valence plus Polarization*),⁽¹²⁸⁾ e incluye el modelo de disolvente continuo COSMO.⁽¹²⁹⁾ Aunque el funcional PBE ha proporcionado resultados razonablemente buenos en una amplia variedad de sistemas ⁽¹³⁰⁾, realizamos un pequeño *benchmarking* para validar nuestro nivel de cálculo. Una vez corregido el error de superposición de base por el método *counterpoise* (CP) y la energía dispersión con la fórmula propuesta por Elstner y col., ⁽¹³¹⁾ los errores asociados a nuestra aproximación resultan ser del orden de décimas de kcal/mol. Por otro lado, no nos sorprende que la inclusión del modelo de disolvente contribuya a un resultado tan bueno. Si se aplica un *cutoff* de más de 8 Å las interacciones que más se ven afectadas son las de naturaleza electrostática (las que tienen mayor alcance), las mismas que se “apantallan” si en el problema se incluye los efectos del disolvente.

Teniendo en cuenta los resultados anteriores, tenemos ciertas garantías a la hora aplicar el método TFEM al modelo POG10 íntegro. Esto se realiza nuevamente sobre 25 estructuras extraídas de la dinámica molecular, utilizando el mismo nivel de cálculo PBE/SVP+COSMO y corrigiendo la dispersión empíricamente. De igual forma, y para poder calcular el cambio de energía asociado a la formación de la triple hélice, se calcula en las mismas condiciones la energía del estado disociado (sin utilizar TFEM). El error de superposición de base, que será intramolecular en el estado disociado, se corrige en ambos casos. El resto de términos necesarios para estimar la energía libre del proceso, se añaden con aproximaciones de naturaleza empírica como MM-PBSA utilizando en general el mismo protocolo que en el primer trabajo del tema 1 (trabajo

³ Sería bueno aclarar que en el caso de la aproximación TFEM y debido a su simplicidad, no se diseñó un programa como tal que pueda ser utilizado por un usuario ajeno a nuestro grupo de investigación. Se utilizaron pequeños scripts que pueden ser reproducidos de un modo sencillo por cualquier potencial usuario que haya leído nuestro manuscrito.

2.2.1). Finalmente, aun obteniendo resultados razonables, se hace una discusión crítica de los valores obtenidos comparándolos con el trabajo 2.2.1. Curiosamente, y a pesar de haber utilizado en parte cálculos QM para la estimación del cambio de energía libre, se tiene un mejor resultado haciendo todos los cálculos a nivel MM.

El último trabajo que presentamos “*Thermochemical Fragment Energy Method for Biomolecules*” es un *benckmarking* más exigente que el propuesto en el trabajo anterior. Se comparan los resultados TFEM con el cálculo QM convencional de todo el modelo POG10 al mismo nivel de teoría anterior sobre cinco estructuras extraídas de la correspondiente dinámica molecular. Se muestra que el nuevo método reduce considerablemente el coste computacional y que, incluso en sistemas con más de 1000 átomos podemos tener errores del orden de las décimas de kcal/mol. En este caso los resultados nuevamente sugieren que la inclusión del modelo de disolvente mejora sustancialmente los resultados. Por otro lado, destacamos que al tratarse de una combinación lineal, este método (TFEM) no sólo nos permitiría el cálculo de energías, sino también de gradientes y segundas derivadas. En cualquier caso, este protocolo debería ser refinado aún más antes de pasar a un uso rutinario del mismo.

2.2.1 Compendio de Publicaciones

En esta sección se recopilan las publicaciones y manuscritos generados en el segundo tema de la Tesis. En ellos se pueden encontrar en detalle los resultados de los estudios introducidos anteriormente, incluyendo toda la información suplementaria.

2.2.1.1 Thermochemical Fragment Energy Method for Biomolecules:

Application to a Collagen Model Peptide

Ernesto Suárez, Natalia Díaz and Dimas Suárez

J. Chem. Theory Comput. **2009**, 5: 1667–1679

Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide

Ernesto Suárez, Natalia Díaz, and Dimas Suárez*

*Departamento de Química Física y Analítica, Universidad de Oviedo,
33006 Oviedo (Asturias), Spain*

Received November 17, 2008

Abstract: Herein, we first review different methodologies that have been proposed for computing the quantum mechanical (QM) energy and other molecular properties of large systems through a linear combination of subsystem (fragment) energies, which can be computed using conventional QM packages. Particularly, we emphasize the similarities among the different methods that can be considered as variants of the multibody expansion technique. Nevertheless, on the basis of thermochemical arguments, we propose yet another variant of the fragment energy methods, which could be useful for, and readily applicable to, biomolecules using either QM or hybrid quantum mechanical/molecular mechanics methods. The proposed computational scheme is applied to investigate the stability of a triple-helical collagen model peptide. To better address the actual applicability of the fragment QM method and to properly compare with experimental data, we compute average energies by carrying out single-point fragment QM calculations on structures generated by a classical molecular dynamics simulation. The QM calculations are done using a density functional level of theory combined with an implicit solvent model. Other free-energy terms such as attractive dispersion interactions or thermal contributions are included using molecular mechanics. The importance of correcting both the intermolecular and intramolecular basis set superposition error (BSSE) in the QM calculations is also discussed in detail. On the basis of the favorable comparison of our fragment-based energies with experimental data and former theoretical results, we conclude that the fragment QM energy strategy could be an interesting addition to the multimethod toolbox for biomolecular simulations in order to investigate those situations (e.g., interactions with metal clusters) that are beyond the range of applicability of common molecular mechanics methods.

Introduction

The idea of representing the total energy of a large molecule as a combination of fragment energies has been considered for decades. To better appreciate their similarities and differences, we will first review several computational approaches for combining fragment energies that have been developed during recent years. We note, however, that other linear-scaling methodologies^{1,2} aimed at construction of the full density matrix of a large system from the fragment density submatrices are beyond the scope of this paper. Thus,

we will discuss first the methods based on the multibody expansion approach and other closely related methods that include implicitly high-order many-body effects into fragment energies using various approximations. We will also comment on the so-called kernel energy method that turns out to be essentially a multibody expansion method. Subsequently, we will review other methods that approximate the quantum mechanical energy of large systems by combining fragment energies on the basis of intuitive and/or thermochemical argumentations. Although we will see that these thermochemically based protocols can be considered as truncated forms of the more general multibody expansion method, they are conceptually simpler and can be readily

* Corresponding author phone: +34-985103689; fax: +34-985103125; e-mail: dimas@uniovi.es.

applicable using many computational tools at a moderate computational cost. In fact, we will formulate yet another variant of the thermochemical fragment energy methods that could be particularly useful to compute the energies of large biomolecular systems. Finally, as a real case application of the proposed method, we will combine fragment-based quantum chemical energies with molecular mechanics and standard quantum chemical calculations in order to compute the relative free energy of the triple-helical form of a collagen model peptide with respect to its monomer state.

Multibody Expansion Method. The so-called cluster expansion method³ has been developed in the framework of solid-state chemistry in order to represent the total energy of an atomic crystal as a linear combination of the characteristic energies of clusters of atoms over a fixed lattice. The coefficients in the cluster expansion are computed using quantum mechanical energy calculations of a few prototype structures. However, the so-constructed functions are not transferable, i.e., they cannot be used for each conceivable configuration of the system. Subsequently, the multibody expansion (MBE) method, also called N -body potentials, or otherwise, cluster potentials, has been developed as a more refined version of the cluster expansion technique.⁴ The MBE method evaluates the total energy as a summation of energies corresponding to isolated atomic clusters extracted from the global structure so that they include systematically two-, three-, and N -body effects. More recently, it has been demonstrated that the MBE approach can be generalized for an *arbitrary* system, whose energy can be uniquely evaluated using series of structure-independent, perfectly transferable, many-body potentials.⁵ In this general MBE formalism, the total energy of an M -particle system (composed of atoms, molecules, or molecular fragments linked covalently) can be expressed as $E_M(A_1, A_2, \dots, A_M)$, where $A_i = \{\mathbf{R}_i, \sigma_i\}$ has the information about the coordinates (\mathbf{R}_i) and the type (σ_i) of the i particle. Since the ordering of the M particles is arbitrary, the functional form of E_M must be such that E_M is invariant to any permutation $A_i \leftrightarrow A_j$.

Representing the total energy by an expansion of a series of N -order (or N -body or N -fragment) energy contributions $E^{(N)}$, we have

$$E_M(A_1, A_2, \dots, A_M) = \sum_{N=1}^M E^{(N)}(A_1, A_2, \dots, A_M) \quad (1)$$

where, in turn, the $E^{(N)}$ terms can be computed from a multiple summation of N -order interaction potentials

$$E^{(N)} = \sum_{m_1 < \dots < m_N}^M V^{(N)}(A_{m_1}, A_{m_2}, \dots, A_{m_N}) \quad (2)$$

where the sum $\sum_{m_1 < \dots < m_N}^M V^{(N)}$ runs over all possible combinations $\{m_1, \dots, m_N\} \in \{1, \dots, M\}$.

Note that eqs 1 and 2 express the total energy E in terms of N -order potentials. In practice, however, one needs to compute the $V^{(N)}$ potentials from energy calculations performed on different subsystems. The general relationship between $V^{(N)}$ and subsystem energies can be obtained through a Möbius inversion as defined in number theory.⁵ The general result is

$$V^{(N)}(A_1, A_2, \dots, A_N) = \sum_{L=1}^N (-1)^{N-L} \sum_{m_1 < \dots < m_L}^N E(A_{m_1}, A_{m_2}, \dots, A_{m_L}) \quad (3)$$

In the above equation, $E(A_{m_1}, A_{m_2}, \dots, A_{m_L})$ stands for the energy of a cluster composed by L fragments labeled by the (m_1, m_2, \dots, m_L) indices. In fact, eq 3 constitutes a unique definition of the N -order interaction potential $V^{(N)}$, which is structure independent because this equation does not carry any information about the environment of the subsystems.⁵ The actual significance of eq 3 can be more easily grasped by deriving the first terms of the N -order expansion leading to the total energy. Thus, the sum of the first-order potentials is just the sum of the energies of the isolated fragments

$$E^{(1)} = \sum_{m_1=1}^M V^{(1)}(A_{m_1}) = \sum_{m_1=1}^M E(A_{m_1}) \quad (4)$$

For the second-order contribution, which can be interpreted as the *excess* energy due to pair interactions, we obtain

$$E^{(2)} = \sum_{m_1 < m_2}^M V^{(2)}(A_{m_1}, A_{m_2}) = \sum_{m_1 < m_2}^M [E(A_{m_1}, A_{m_2}) - E(A_{m_1}) - E(A_{m_2})] \quad (5)$$

and, of course, $E_M \approx E^{(1)} + E^{(2)}$ defines the well-known pairwise additive approximation to the total energy. Analogously, the three-body $E^{(3)}$ contribution, which collects the $V^{(3)}$ potentials, is the additional energy due to three-body effects, and that cannot be assessed from a two-body representation

$$V^{(3)} = \sum_{m_1 < m_2 < m_3}^M [E(A_{m_1}, A_{m_2}, A_{m_3}) - E(A_{m_1}) - E(A_{m_2}) - E(A_{m_3}) - V^{(2)}(A_{m_1}, A_{m_2}) - V^{(2)}(A_{m_1}, A_{m_3}) - V^{(2)}(A_{m_2}, A_{m_3})] \quad (6)$$

Finally, it may be interesting to note that the MBE equations can be rewritten in terms of the so-called mutual information functions (MIFs),⁶ which have been used to compute the configurational entropy of flexible molecules. Thus, the MIF expansion approaches the full-dimensional configurational probability distribution by including systematically N -order correlations among the internal degrees of freedom; likewise, the successive $V^{(N)}$ potentials include the N -order effects on the total energy. Similarly, the energy of a system composed of M arbitrary fragments can be expanded using the MIFs in the following form

$$E_M(A_1, A_2, \dots, A_M) = \sum_{i=1}^M E(A_i) - \sum_{m_1 < m_2}^M I_2(A_{m_1}, A_{m_2}) + \dots + (-1)^{N-1} \sum_{m_1 < \dots < m_N}^M I_N(A_{m_1}, \dots, A_{m_N}) \quad (7)$$

where the mutual information function $I_N(A_{m_1}, \dots, A_{m_N})$ combines the energies of all the clusters formed by N fragments

$$I_N(A_{m_1}, \dots, A_{m_N}) = \sum_{L=1}^N (-1)^{L+1} \sum_{m_1 < \dots < m_L} E(A_{m_1}, \dots, A_{m_L}) \quad (8)$$

Note that the mathematical form of the MBE and MIF expressions are identical due to the fact that $(-1)^{N-L} \equiv (-1)^{N+L}$.

Kernel Energy Method is an MBE Method. At this point, it is convenient to simplify the notation used in the MBE equations by replacing $E(A_{m_1}, A_{m_2}, \dots, A_{m_L})$ (the energy of the subsystem with L fragments) with $E_{ijk\dots}$ (the energy of the subsystem composed of the i, j, k, \dots particles or fragments). In this way, the pairwise additive approximation for a system composed of a total of M fragments can be written as

$$E_M = \sum_{i=1}^M E_i + \sum_{i=1}^M \sum_{j=i+1}^M (E_{ij} - E_i - E_j) \quad (9)$$

In recent years, the so-called kernel energy method (KEM) has been utilized to compute the quantum mechanical (QM) energy of large biomolecules^{7–11} by representing a full molecule by smaller *kernels* of atoms (i.e., fragments A_i). The majority of the KEM applications that have been reported to date compute the total energy “by summation over the energy contributions of all *double* kernels reduced by those of any *single* kernels, which have been overcounted in the sum over double kernels”,⁸ that is, by means of the following expression

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M-2) \sum_{i=1}^M E_i \quad (10)$$

However, it can be easily demonstrated (see Supporting Information) that the original KEM energy formula is equivalent to the MBE pairwise additive approximation.

Several KEM applications on biomolecules have been reported in which the dangling bonds of the molecular fragments are saturated with hydrogen atoms before carrying out the corresponding fragment energy calculations. However, the presence of the H-link atoms introduces an error in the computation of the total energy given that the validity of the MBE equations requires that only the actual fragments are considered in the calculations. Nevertheless, if the fragments are large enough and the total number of fragments is relatively low, the associated error can be reasonably small. Of course, the H-link error can be further reduced by including higher order MBE terms given that these terms progressively account for the environment of each fragment by considering larger and larger clusters of fragments. This has been done in a recent article in which the KEM equation is expanded up to fourth order¹¹ through a cumbersome derivation that follows an MBE recipe employed in a former study of water clusters.¹²

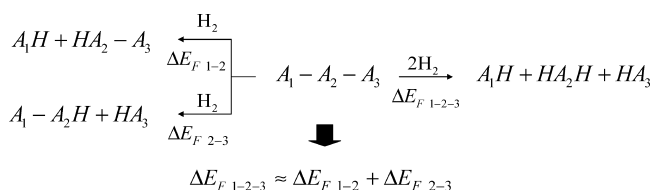
Electrostatically Embedded MBE Methods. In principle, the pairwise additive approximation defined by eq 9 is not enough to accurately compute the total energy of complex systems. Unfortunately, the calculation of higher order MBE terms is extremely expensive in terms of computer time. In order to overcome the limitations of second-order methodologies at a reasonable computational cost, some authors

proposed to compute the energies of the individual fragments (E_i) and fragment pairs (E_{ij}) taking into account the electrostatic field of the rest of the system.^{13–18} For example, in the fragment molecular orbital (FMO) method, the energies of the different fragments are computed by iteratively solving *effective* fragment Hamiltonians that include the electrostatic effects from the electrons in the surrounding ($M-1$) fragments as well as from all nuclei in the total molecule.^{14,19} The resulting FMO energies are then combined using MBE equations of order 2 or 3 to derive the total energy. A similar alternative for noncovalently connected fragments is the electrostatically embedded many-body expansion (EE-MBE).^{16–18} The energy of each cluster is calculated in the presence of the electric field due to the fixed partial atomic charges of the surrounding fragments. A significant improvement in the electrostatically embedded second- and third-order energies for a series of water clusters is found when compared with the results of standard MBE calculations.¹⁶

Molecular Tailoring Approach. The so-called molecular tailoring approach (MTA)²⁰ divides the total system into *overlapping* fragments and subsequently estimates the total energy by summing the fragment contributions and then subtracting the energies of fragment *intersections*. This means that interactions between nonoverlapping fragments are neglected in the MTA method and that each fragment intersection formally accounts for N -body effects to the total energy, with N being the number of overlapping fragments at the particular intersection. This strategy is somehow equivalent to employing localized multibody expansions, and therefore, the MTA approach can be considered as a *flexible* MBE method. The MTA method can also compute one-electron properties of the full system by combining the fragment density matrices into a single density matrix for the whole system.²¹

Molecular Fractionation with Conjugate Caps. The so-called molecular fractionation with conjugate caps (MFCC) scheme also estimates the total energy of large systems from calculations performed on fragments. The MFCC method was originally designed to compute the QM interaction energy between a protein and a small ligand,²² but this method has been expanded to predict the total energy of protein molecules.²³ In this approach, the protein is divided into fragments $A_i = (-C_\alpha HR_i - CO - N_{i+1} H -)$, with R_i being the side chain of the i amino acid residue and N_{i+1} is the backbone N atom of the $(i+1)$ amino acid. Instead of H-link atoms, two “conjugate caps”, NH_2- and $-C_\alpha H_2 R_{i+1}$, are placed at the corresponding $C_{\alpha,i}/N_{i+1}$ atoms to saturate the exposed valence sites of each fragment A_i . The total energy of an M -residue protein molecule is first approximated by summing the energies of the (capped) fragments and then subtracting the energies of the $NH_2-C_\alpha H_2 R_{i+1}$ conjugate caps. This first-order approximation is then corrected ad hoc by adding a second-order term ($\delta E^{(2)}$) that accounts for the pairwise interaction energy between non-neighboring fragments. The final MFCC expression is

$$E_M = [E(A_1 - C_\alpha H_2 R_2) + \sum_{i=2}^{M-1} E(NH_2 - A_i - C_\alpha H_2 R_{i+1}) +$$

Scheme 1

$$(11) \quad E(\text{NH}_2 - A_M)] - \left[\sum_{i=1}^{M-1} E(\text{NH}_2 - C_\alpha \text{H}_2 \text{R}_{i+1}) \right] + \delta E^{(2)}$$

To compute the $\delta E^{(2)}$ contribution, the fragments are capped with H-link atoms as in the KEM scheme. Alternatively, another variant of the MFCC method has been proposed that uses only fragment energies, which are computed in the presence of the electrostatic field created by point charges representing the non-neighboring residues.²⁴

Systematic Molecular Fragmentation. As we will see later, the MFCC expression¹¹ can be justified by means of simple thermochemical arguments on the basis of formal fragmentation processes of the protein system. In fact, the thermochemical approach for computing the fragment-based energy of large molecules has already been explored systematically by Collins et al.²⁵ The basic reasoning behind the generalization proposed by Collins et al. is summarized in Scheme 1, which shows a generic molecular system composed of three fragments ($A_1-A_2-A_3$) that can be formally broken through three different fragmentation processes.

The key approximation in the protocol of Collins et al. is that the reaction energy for the total fragmentation of $A_1-A_2-A_3$ (ΔE_{F1-2-3}) is estimated as the sum of the reaction energies corresponding to the two single-fragmentation processes (i.e., $\Delta E_{F1-2} + \Delta E_{F2-3}$). The straightforward consequence of this approximation is that the energy of the total system can be expressed as a combination of the energies of the three smaller subsystems

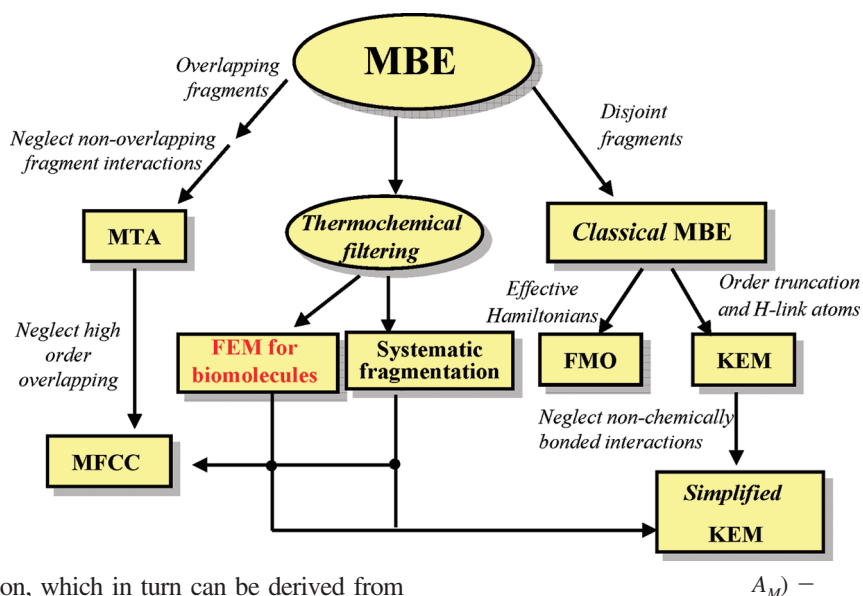
$$E_{123} = E_{12} + E_{23} - E_2 \quad (12)$$

In principle, Collins et al. employ both chemical topology and computer cost considerations in order to choose the best site at which a large molecule is cut so that the resulting A_2 fragment is (a) large enough to reasonably neglect the interaction between A_1 and A_3 and (b) simultaneously small enough to compute the energy of the A_1-A_2H fragment using high-level QM methods. If the accompanying HA_2-A_3 fragment is too large, the fragmentation protocol defined in Scheme 1 is then applied iteratively until all the produced fragments can be described quantum mechanically. Ultimately, this thermochemical approach results in the total energy being approximated by a linear combination of fragment energies, whose precise form depends on the nature of the chemical system and on the chemical topology and computer cost considerations. Like in the MFCC method, the systematic fragmentation technique can be augmented with a nonbonded energy correction by computing the interaction energy between two nonchemically bonded fragments if their separation is below a certain threshold.²⁵

Comparison of the Different Methods. Although largely unnoticed in some of the previous works, the MBE formalism provides the general framework for developing computational strategies aimed at the evaluation of the total energy of large systems from subsystem (fragment) energies (see Scheme 2). Thus, the FMO method, the various KEM formulas, and the MFCC expression with pairwise interactions can be classified as MBE techniques that include N -body effects through fragment energy calculations. Similarly, the systematic fragmentation method of Collins et al. can be generated directly from the MBE expansion by neglecting all the MBE interaction potentials beyond second order and using an additional chemical topology criterion to neglect a large number of second-order contributions. We can also see in Scheme 2 that inclusion of the H-link atoms to cap the exposed valence sites of the fragments extracted from a covalent system makes the Collins' fragmentation method nearly identical to the simplified version of the KEM method in which only the chemically bonded *double kernels* are considered.⁸ Thus, once a fragmentation scheme has been applied, the same energy terms are actually computed in the two methods. Similarly, the systematic fragmentation proposed by Collins et al. encompasses the effective MFCC in which only fragment energies are considered. On the other hand, the MFCC method can be considered as a particular case of the MTA formalism given that the MFCC-capped fragments are equivalent to the MTA overlapping fragments and the MFCC conjugate caps would correspond to fragment intersections in the MTA approach. However, while the MFCC fragments are built to make simple overlaps (i.e., each atom can only be part of one or two fragments), the MTA method admits more complex fragment overlaps among N fragments. These and other interrelationships show that in general fragment energy methods assume a similar *ansatz*.

Goals of the Present Work. In principle, the ability to perform on a routine basis fragment energy calculations on large biomolecules could be very useful to predict their energetic properties using high-level QM methodologies. Fortunately, previous test applications have shown that high-order MBE contributions contain many more energetic terms than those that are actually required to derive the total energy from fragment energies within a reasonable accuracy. In this way and taking into account that proteins and nucleic acids are linear polymers that exhibit many repetitive secondary structural motifs, we believe that a thermochemical approach complemented with a distance-based criterion is probably the best option to formulate a linear scaling fragment-based energy method for biological molecules. This approach, which can be considered as a thermochemical truncation of the multibody expansion, is also computationally advantageous because the required energetic terms can be easily computed using standard methodologies. Another advantage of the thermochemical framework is that the successive fragmentation energies involved in the formal degradation of the biomolecule can be computed taking into account the effect of a solvent continuum in the QM Hamiltonian. Thus, in this work, assuming a simple fragmentation process, we will derive a fragment energy formula for estimating the total energy of a biomolecule as function of a cutoff criterion. On one hand, we will show that our fragment energy method (FEM in Scheme 2) can have a broader applicability

Scheme 2

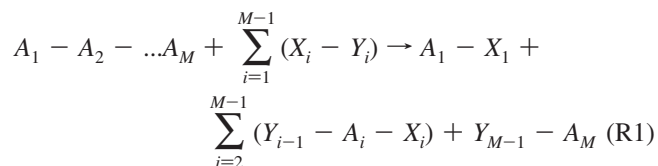


than the MFCC equation, which in turn can be derived from our approach as a particular case. On the other hand, with respect to the more general thermochemical scheme of Collins et al., our expression will be more readily applicable to (and limited to) large biomolecular systems in which a natural choice for the formal fragmentation processes can be easily made. In addition, more emphasis will be placed upon the consistent use of a cutoff criterion in the fragment energy calculations, the inclusion of solvent effects, the mixing of QM and molecular mechanical calculations, and the potential implementation of the fragment-based energy methods within the context of QM/MM methodologies.

Theory

For the sake of simplicity, we will consider a macromolecule \mathbf{P} that is a linear chain of M fragments A_i interconnected through covalent bonds ($A_1-A_2-\dots-A_M$). For example, if \mathbf{P} is a protein, A_i could be a single amino acid or a secondary structure element. We do note, however, that the same equations based on fragment energies would result for more complex topological patterns connecting the A_i fragments like in cyclic or branched macromolecules.

The total fragmentation of \mathbf{P} can be achieved through the following formal reaction



Note that every fragment linkage in the \mathbf{P} molecule is broken through insertion of a specific X_i-Y_i molecule(s) into the A_i-A_{i+1} bond. If \mathbf{P} is not a linear chain, then X_i and Y_i would stand for all the molecular caps that are required to saturate the exposed bonds after having removed the A_i fragment from the rest of the \mathbf{P} molecule. In any case, the total energy change corresponding to the above formal reaction is

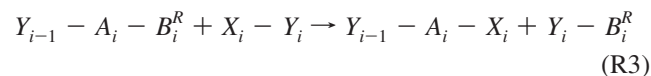
$$\Delta E = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + E(Y_{M-1} -$$

$$A_M) - \sum_{i=1}^{M-1} E(X_i - Y_i) - E(\mathbf{P}) \quad (13)$$

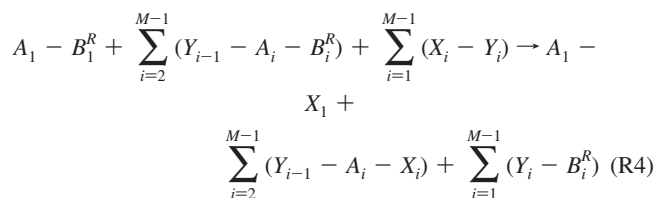
The thermochemical approximation to compute ΔE can be introduced as follows: we compute first the reaction energy for the fragmentation step in which the A_1 fragment is removed. However, we assume that the reactants involved in the first fragmentation process are subsystems of \mathbf{P} that are defined on the basis of some geometric and/or chemical-structure criterion. The same criterion, denoted onward as the R criterion, should be applied consistently along the \mathbf{P} backbone structure. Perhaps the simplest criterion for defining the reactants could be to impose a layer cutoff around the leaving A_1 fragment, but other choices like sequence proximity could be used. Thus, assuming that a well-defined R criterion is used, the first fragmentation reaction can be written as



where B_1^R represents a buffer region, which includes all the neighboring atoms (or fragments A_i) that are around A_1 in the \mathbf{P} structure depending on the R criterion being used. Similarly, the fragmentation process for the A_i-A_{i+1} bond can be represented by the following chemical equation



where the closer atoms or fragments around A_i excepting those in $A_{i-1}, A_{i-2}, \dots, A_1$ are included in the buffer B_i^R . The sum of the $M-1$ fragmentation processes defined in this manner leads to the following chemical equation



In this way, the energy change for the total fragmentation of \mathbf{P} through the R -dependent fragmentation processes (ΔE^R) is given by

$$\Delta E^R = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1} E(Y_i - B_i^R) - [E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1} E(X_i - Y_i)] \quad (14)$$

Extracting the exact fragmentation energy ΔE from eq 13 and defining $\delta E = \Delta E^R - \Delta E$, we can combine eqs 13 and 14 in order to exactly express the total energy of the system $E(\mathbf{P})$ in terms of the fragment energies and the δE difference

$$E(\mathbf{P}) = \left[E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + E(Y_{M-1} - A_M) \right] - \left[\sum_{i=1}^{M-1} E(Y_i - B_i^R) \right] + \delta E(\mathbf{B}^R, \mathbf{Y}) \quad (15)$$

where the δE difference is expressed as a function of $\mathbf{B}^R = \{B_i^R\}$ and $\mathbf{Y} = \{Y_i\}$. This is a consequence of the fact that $E(\mathbf{P})$ is rigorously independent of \mathbf{B}^R , $\mathbf{X} = \{X_i\}$, and \mathbf{Y} and that the terms in the square brackets are independent of \mathbf{X} (i.e., the identity of the X_i moieties is irrelevant).

For practical applications of the thermochemical fragment energy eq 15, the δE term must be neglected. To increase the accuracy of the fragment-based energy calculations, one straightforward solution would be to systematically increase the R criterion in order to include larger portions of the remaining \mathbf{P} molecule in the B_i^R buffer regions until reaching a reasonable compromise between accuracy and computational cost. The best systems for which we can efficiently apply this simple strategy would be *linear* structures like carbon nanotubes, DNA segments, collagen molecules, etc. Of course, in the case of more compact systems like globular proteins, a larger computational cost and a lower accuracy can be expected for the same R criterion because the buffer regions would contain many more atoms and truncation effects would be more important. However, we could also use the well-known QM/MM methodologies in order to calculate the reaction energies of the fragmentation steps using the same settings as those that are typically employed in routine QM/MM calculations. In this case, the R criterion would be applied to select the size of the QM region while the rest of the system would be treated classically. Thus, like in the electrostatically embedded variants of the MBE methodologies, we expect that QM/MM calculations of fragmentation energies could account for high-order effects within the thermochemical approach.

As above mentioned, we can particularize the general eq 15 to obtain the MFCC equation for a protein system. This can be done by matching Y_i by $-\text{NH}_2$ and B_i^R by $-\text{R}_{i+1}\text{C}_\alpha\text{H}_2$, which are the “conjugate caps” adopted in the MFCC scheme. In our thermochemical terminology, these choices are equivalent to consider $X_i-\text{NH}_2$ as the capping dimers as well as to adopt a minimum sequence proximity R criterion for defining the B_i^R groups. Then eq 15 becomes

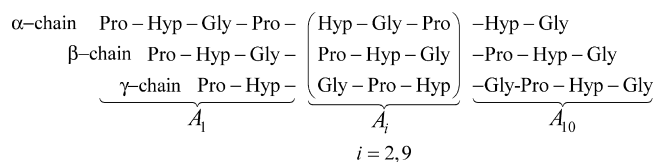
$$E(\mathbf{P}) = [E(A_1 - \text{C}_\alpha\text{H}_2\text{R}_2) + \sum_{i=2}^{M-1} E(\text{NH}_2 - A_i - \text{C}_\alpha\text{H}_2\text{R}_{i+1}) + E(\text{NH}_2 - A_M)] - [\sum_{i=1}^{M-1} E(\text{NH}_2 - \text{C}_\alpha\text{H}_2\text{R}_{i+1})] + \delta E \quad (16)$$

If we compare this equation with eq 11, we see that the “non-neighboring interactions” ($\delta E^{(2)}$) in the MFCC approach²³ constitutes an approximation to the actual error (δE) committed in the calculation of the global fragmentation energy. We note in passing that the same energy contributions collected in eq 16 can be associated to other formal fragmentation processes by changing accordingly the definition of the A_i fragments and the corresponding conjugated caps. For example, expression 16 also results if the A_i fragment corresponds to the i residue and $Y_i = \text{H}$.

Finally, it may be interesting to note that our approach, like with all the MBE-like methods, computes the total energy as a linear combination of fragment energies. As gradient is a linear operator, its application over the fragment energy expression would be straightforward as previously noticed in other works.^{20,25} In this way, both energy and gradient values for the total system could be obtained from fragment calculations using similar approximations and techniques as those typically used by the QM/MM methodologies.^{26,27}

Results and Discussion

In many of the previous works, the viability of fragment-based energy methods has been assessed by means of proof of principle applications, that is, by carrying out single-point calculations and using relatively low QM levels of theory. However, most of the biomolecules are flexible molecular systems in aqueous solution, and therefore, in actual applications, structures for performing fragment-based QM calculations should be provided by Monte Carlo or molecular dynamics (MD) simulations using either explicit or implicit solvent models. In this respect, we think that classical MD simulations still constitute the most reasonable alternative to generate the biomolecular structures for the subsequent fragment QM calculations. This approach would be similar to the molecular mechanics Poisson–Boltzmann method,²⁸ which predicts mean values of free energies of biomolecules in solution as estimated over a series of representative snapshots extracted from classical MD simulations. Moreover, we also note that various levels of approximation could be required in the fragment energy calculations. For example, a standard density functional level of theory combined with an implicit solvent model can take into account both the intramolecular electronic effects and the solute–solvent electrostatic interactions. Other free-energy terms such as attractive dispersion interactions or thermal contributions could be calculated using molecular mechanics (MM). We believe that this and other technical issues like the counterpoise correction of the basis set superposition error (BSSE) in the QM calculations should be explicitly considered in the test calculations in order to assess the actual performance of the fragment QM energy calculations in the context of multimethod approaches to simulating biomolecules. There-

Scheme 3

fore, we decided to reexamine in this work the problem of the stability of triple-helical collagen model peptides by combining our fragment energy expression with previous MD and MM data that have been reported by us recently.²⁹

Many collagen model peptides with 30–45 amino acids have been synthesized to investigate the thermal stability and folding of the triple-helix domain of natural collagen. These peptides, which are also known as triple-helical peptides (THPs), assemble spontaneously to form a triple-helix complex that can be characterized using a wide array of experimental techniques.³⁰ The THP molecules present a characteristic triple-helix structure composed of three peptide chains, each in an extended, left-handed polyproline II-like helix, which are staggered by one residue and then supercoiled about a common axis in a right-handed manner. The close packing of the three chains requires the presence of a sterically small glycine residue at every third position. The test calculations reported in this work were performed on the prototypical [(Pro-Hyp-Gly)₁₀]₃ system (labeled as **POG10**), which contains many proline and 4(*R*)-hydroxyproline (Hyp) residues that largely stabilize the triple-helix conformation.^{31,32}

Selection of a Fragmentation Process. The collagen model for our test calculations, **POG10**, contains three peptide chains (labeled α , β , and γ) with 30 amino acids per chain. As mentioned above, the fragment energy expression, eq 15, that has been derived by assuming that the **P** macromolecule is a linear chain, is also applicable for more complex macromolecules like **POG10**. To this end, we describe the triple helix as a *linear* arrangement of 10 fragments comprising each of three *triplets* of residues from the α , β , and γ chains (see Scheme 3). The resulting building blocks or fragments A_i will be termed as *triplets*. A pair of consecutive *triplets*, A_i – A_j , is interconnected through three peptide linkages corresponding to the α , β , and γ chains. We chose this mode of partitioning because it minimizes the interactions between nonconsecutive *triplets* and maximizes the number of interactions among the three peptide chains within each *triplet*.

After having chosen a structurally and computationally convenient partitioning of **POG10**, we can define more precisely the formal fragmentation processes required for the fragment-energy calculations based on eq 15. More specifically, we see in Figure 1 how the terminating Y_i group attached to the *N*-terminal end of the A_i triplet comprises three acetyl groups for the α , β , and γ peptide chains, whose coordinates are extracted from the *C* end of the previous A_{i-1} triplet and augmented with the required H-link atoms. Similarly, the buffer group B_i^R attached to the *C* end of the A_i triplet includes the adjacent A_{i+1} triplet plus three *N*-methyl moieties extracted from the A_{i+2} fragment (this choice of B_i^R is equivalent to a ~ 9 Å cutoff around the leaving A_i

fragment). This formal fragmentation process can also be applied straightforwardly to obtain the energy of the individual peptide chains α , β , and γ . In this case, the corresponding A_i , B_i^R , and Y_i groups include residues located in the same chain.

Comparison between Conventional and Fragment-Based QM Energies. Before computing the energy of the full **POG10** system, we assessed the combined quality of the fragment energy calculations and the collagen partitioning in order to reproduce the energetic properties of a relatively large collagen subsystem. The size of the selected subsystem, [Ace-(Pro-Hyp-Gly)₄-Nme]₃ (456 atoms), still allowed us to carry out full QM calculations. Following similar prescriptions to those represented in Figure 1, four different fragments (A_i) can be distinguished in this model. We computed both the interaction energy among the three peptide chains and the absolute energy of the THP model. The calculations were performed on 25 structures that were built using the coordinates of the central region of POG10 extracted from MD snapshots (see Table S1 in the Supporting Information).²⁹ As described in the Computational Section, the energy calculations were carried out using a density functional level of theory (PBE/SVP) combined with the COSMO solvent model. The intramolecular dispersion energy is included via an empirical method. The BSSE arising from the interchain interactions is corrected using the standard counterpoise (CP) method. In the case of the fragment energy calculations, the CP correction was applied to the fragment electronic energies, that is, the electronic energies of the A_1 – B_1^R , Y_1 – A_2 – B_2^R , ..., fragments extracted from one peptide chain (e.g., α) were computed in the presence of the ghost basis functions located in the equivalent fragments from the other two chains (e.g., β and γ). For the full QM calculations, the CP recipe was used to correct the BSSE of the electronic energies of the full peptide chains.

The total interaction energy of [Ace-(Pro-Hyp-Gly)₄-Nme]₃ can be estimated from the combination of five energy terms using eq 15 (see Table 1). Similarly, the energy of each Ace-(Pro-Hyp-Gly)₄-Nme peptide chain can be computed from the corresponding fragment energies. In this way, we derived an average interaction energy (ΔE_{int}) of -29.4 ± 0.2 kcal/mol that matches perfectly the *exact* value (-29.5 ± 0.2 kcal/mol) according to conventional QM calculations.

Since ΔE_{int} is a relative quantity, it can be expected that the fragment energy calculations would benefit from partial cancelation of errors. However, we see in Table 1 that the total energy E of the whole system in aqueous solution can be computed accurately using the fragment energies given that the error in the mean value of the fragment-based energies with respect to the exact full QM value is rather small, 0.0001 au (~ 0.1 kcal/mol). Table S1 (Supporting Information) shows that small errors arise also in each of the individual structures considered in the calculations. We also see in Table 1 that the observed accuracy in the total energy benefits from a partial cancelation of errors in the computation of the individual energetic components, which result in energy differences of +1.4 (gas-phase energy) and -1.5 kcal/mol (solvation energy) between the fragment-based and the exact values. Although the accuracy in the gas-phase

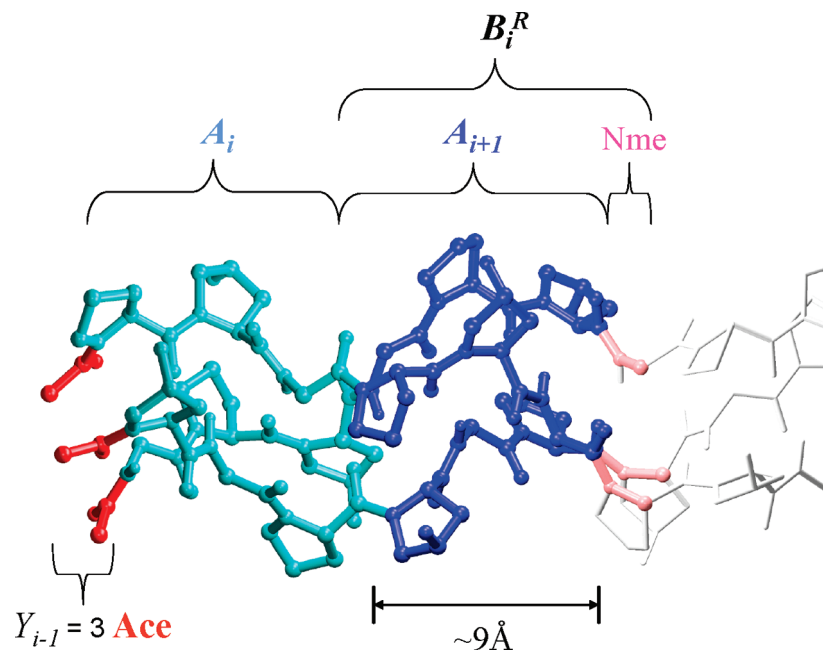


Figure 1. Ball and stick model of the **POG10** triple helix. The various moieties of **POG10** involved in the formal i -fragmentation step ($i \geq 2$) are shown in different colors. See text for details.

Table 1. Average Values and Standard Deviations of the Interchain Interaction Energies (ΔE_{int} , in kcal/mol of peptide) for the [Ace-(Pro-Hyp-Gly)₄-Nme]₃ System^a

	$A_1-B_1^R$	$Y_1-A_2-B_2^R$	$Y_2-A_3-B_3^R$	$Y_1-B_1^R$	$Y_2-B_2^R$	[Ace-(Pro-Hyp-Gly) ₄ -Nme] ₃		$\Delta_{\text{FRAG-CONV}}$
						FRAG	CONV	
$\Delta \bar{E}_{\text{int}}$	-14.9 ± 0.1	-15.1 ± 0.1	-15.2 ± 0.1	-7.8 ± 0.1	-7.9 ± 0.1	-29.4 ± 0.2	-29.5 ± 0.2	0.1
\bar{E}	-6329.7247 (0.0025)	-6329.7250 (0.0021)	-6329.7254 (0.0021)	-3536.8993 (0.0016)	-3536.8983 (0.0014)	-11915.3775 (0.0032)	-11915.3776 (0.0032)	0.1
\bar{E}_{gas}	-6329.5131 (0.0023)	-6329.5137 (0.0023)	-6329.5131 (0.0022)	-3536.7781 (0.0015)	-3536.7769 (0.0015)	-11914.9849 (0.0032)	-11914.9872 (0.0033)	1.4
$\Delta \bar{G}_{\text{COSMO}}^{\text{elec}}$	-87.1 (0.2)	-87.2 (0.3)	-87.7 (0.3)	-54.8 (0.2)	-55.4 (0.3)	-151.9 (0.4)	-150.4 (0.4)	-1.5
\bar{E}_{disp}	-105.3 (0.3)	-105.4 (0.4)	-105.3 (0.3)	-53.7 (0.2)	-53.8 (0.2)	-208.4 (0.5)	-208.5 (0.5)	0.1

^a Average values and standard errors (in parentheses) of the various energy components for the THP fragments: total energy in solution, E , in au; gas-phase energy, E_{gas} , in au; electrostatic solvation energy, $\Delta \bar{G}_{\text{COSMO}}^{\text{elec}}$, in kcal/mol; and empirical dispersion energy, E_{disp} , in kcal/mol. Mean values of the total energies as obtained with the fragment-based (FRAG) and conventional (CONV) calculations and their differences ($\Delta_{\text{FRAG-CONV}}$, in kcal/mol) are also indicated.

Table 2. Average Values (kcal/mol of peptide) for the Different Energy Components of the Interaction Energy among the **POG10** Peptide Chains^a

$\Delta E_{\text{PBE/SVP}}^{\text{CP-uncorrected}}$	BSSE	$\Delta \bar{G}_{\text{COSMO}}^{\text{elec}}$	$\Delta \bar{E}_{\text{disp}}$	$\Delta \bar{E}_{\text{int}}^b$
-105.6 (1.1)	85.7 (0.2)	37.1 (1.0)	-82.5 (0.1)	-65.4 (0.2)

^a Standard errors are given in parentheses. ^b $\Delta \bar{E}_{\text{int}} = \Delta E_{\text{PBE/SVP}}^{\text{CP-uncorrected}} + \text{BSSE} + \Delta \bar{G}_{\text{COSMO}}^{\text{elec}} + \Delta \bar{E}_{\text{disp}}$.

energy (~ 0.002 au) is comparable to that reported in previous fragment energy calculations,^{10,14,20} these results suggest that inclusion of solvent effects in the fragment QM calculations should improve the accuracy of the fragment-based approaches given that the intramolecular long-range interactions could be dampened out by the electrostatic screening exerted by the surrounding solvent continuum.

Due to the linear structure of collagen, we expect that the performance of the fragment-energy calculations for larger collagen models would be equally satisfactory and that other molecular properties of collagen molecules (e.g., gradients) could be also computed within a reasonable accuracy. Finally, we note that, in terms of CPU time, a single-point energy calculation on the [Ace-(Pro-Hyp-Gly)₄-Nme]₃ system using the fragment approach took about 9 h on one x86-64

processor. The same energy value obtained with conventional QM calculations required about 80 h of CPU time.

Fragment Calculations on the POG10 Triple Helix. The results of our fragment energy calculations on the full **POG10** system (1089 atoms) are summarized in Table 2, which contains the average values of the various energetic components contributing to the interchain interaction energy. The calculations were done on 100 snapshots extracted from our previous MD simulation.²⁹ The total interaction energy amounts to -65.4 kcal/mol of peptide, which gives an average value of -6.5 kcal/mol for every $-(\text{Pro-Hyp-Gly})$ -triplet of residues. As expected, all the energy components considered in the calculations (gas-phase electronic energy, empirical dispersion energy, and electrostatic solvation

energy) contribute significantly to the interaction energy. Of particular interest can be the large weight of the BSSE as estimated by the CP calculations, 85.7 kcal/mol. Clearly, the omission of the BSSE corrections would have resulted in an unphysical overestimation of the interaction energy. On the other hand, the inability of the PBE DFT functional to recover most of the intermolecular dispersion energy justifies the addition of the empirical dispersion energy. In fact, the combination of DFT QM methods and empirical dispersion energy has been used in previous computational studies that apply DFT to study weak nonpolar interactions.^{33–35} Although the three peptide chains intertwined into the triple helix establish many hydrogen-bond interactions that can be described reasonably by the PBE calculations, we see in Table 2 that the dispersion energy is the largest stabilizing contribution to the interchain interaction energy of the **POG10** triple helix. Hence, it turns out that the close packing of the peptide chains plays a crucial role in the overall stabilization of the triple helix.

Perhaps the bottom line from the calculations summarized in Table 2 is that the QM fragment energy approach may constitute a promising alternative for studying the intermolecular interactions in large biomolecules. For the collagen model peptide studied in this work, the error introduced by the fragmentation technique can be rather small (<1 kcal/mol) as suggested by the preliminary test calculations. However, we do note again that when using a DFT level of theory in the fragment calculations for large biomolecules, correction of the BSSE and inclusion of dispersion energy are a must in order to obtain meaningful results for interaction energies.

Intramolecular BSSE. As shown in Table 2, the CP correction to the interchain interaction energy is quite large, +85.7 kcal/mol at the PVE/SVP level, due to the large size of the **POG10** system and the relatively small size of the double- ζ SVP basis set. In principle, the use of larger basis sets should reduce significantly the magnitude of the BSSE but at the cost of increasing the CPU time. Nevertheless, it is most likely that assessing and correcting the BSSE will also be required when carrying out fragment energy calculations on biomolecules using medium-sized basis sets (cc-pVDZ, TZVP, ...). Moreover, it is becoming increasingly clear that the relative energies of different conformations of large and flexible biomolecules are quite sensitive to the size of the basis set and that part of this dependence arises from the *intramolecular* BSSE.³⁶ Although this (presumably small) effect has been commonly ignored so far, there is now some solid computational evidence in the recent literature indicating that the intramolecular BSSE can severely impair the accuracy of the energetic QM predictions for polypeptide systems.^{36–38}

Given that we are interested in computing the relative stability of the triple-helix conformation with respect to the compact form of the isolated chains (see below), we decided to estimate the magnitude of the intramolecular BSSE in our QM calculations. For this purpose, the CP method of Boys and Bernardi could be applied by taking atomic fragments, but this alternative would result in a large number of extra QM calculations as well as in problems in the assignment

of charge, multiplicity, and electronic state of the atomic fragments.³⁹ Hence, we followed a more pragmatic approach that consists of the definition of proper molecular fragments within the large system and adding H-link atoms to saturate the exposed chemical bonds. Subsequently, the BSSE in the interaction among the resulting fragments is computed using the standard CP procedure. A similar approach has been employed previously by other authors.³⁶ For example, Valdés et al. estimated the intramolecular BSSE in [*n*]-helicene molecules consisting of all-ortho-annulated benzene rings by computing the CP-corrected interaction energies of benzene pairs, in which the Cartesian coordinates of the C atoms are identical to those in the helicene.³⁶

After some computational experimentation, we decided to employ the following fragmentation protocol for estimating the intramolecular BSSE of the **POG10** peptide chains. (1) For each **POG10** structure, a pair list of nonbonded (beyond 1–4) interactions involving heavy atoms is built using a distance criteria ($X \cdots Y < 4.0 \text{ \AA}$). (2) Each peptide chain is broken into four smaller fragments by removing three glycine residues. These glycine residues are automatically selected in order to *maximize* the number of nonbonded interactions among the resulting fragments (see Figure 2a and 2b). H-Link atoms are added to saturate the exposed bonds. (3) The standard CP method is used to compute the value of the BSSE corresponding to the interactions among the four fragments (*intra*-BSSE₁; see Figure 2b). (4) The BSSE due to the interactions between the formerly removed glycine residues and the nearby groups is estimated by building a molecular cluster in which the three glycine residues are surrounded by the closer residues. Then the CP procedure is applied again to estimate the BSSE arising from the simultaneous interactions between the three glycines and the rest of the groups (*intra*-BSSE₂; see Figure 2c). (5) The total intramolecular BSSE of the peptide chain is approximated by adding together the two BSSE values computed in 3 and 4.

The QM calculations for estimating the intramolecular BSSE were done on 100 MD snapshots of the free **POG10** chain.²⁹ Thus, we found that, at the PBE/SVP level, the average value of the intramolecular BSSE for the isolated **POG10** chain in its folded state amounts to 92.7 kcal/mol of peptide, which is even greater than the BSSE related to the interchain interactions in the triple-helix state (85.7 kcal/mol). For the sake of consistency, the same protocol was applied on each of the three chains in the triple-helix conformation. In this case, the peptide chains are quite extended and their intramolecular BSSE is predicted to be only 3.1 kcal/mol on average. All these CP-corrected QM calculations can be combined to estimate the energetic penalty for the folded **POG10** chain to adopt its extended conformation in the triple helix, the average value being +30.8 kcal/mol (in terms of $E_{\text{PBE/SVP}} + \text{BSSE}_{\text{intra}} + E_{\text{disp}} + \Delta G_{\text{COSMO}}^{\text{elec}}$). Neglecting the intramolecular BSSE in the folded state of **POG10** would lead to a very large unrealistic value (~120 kcal/mol) for the relative energy between the folded and the extended forms of the peptide chain.

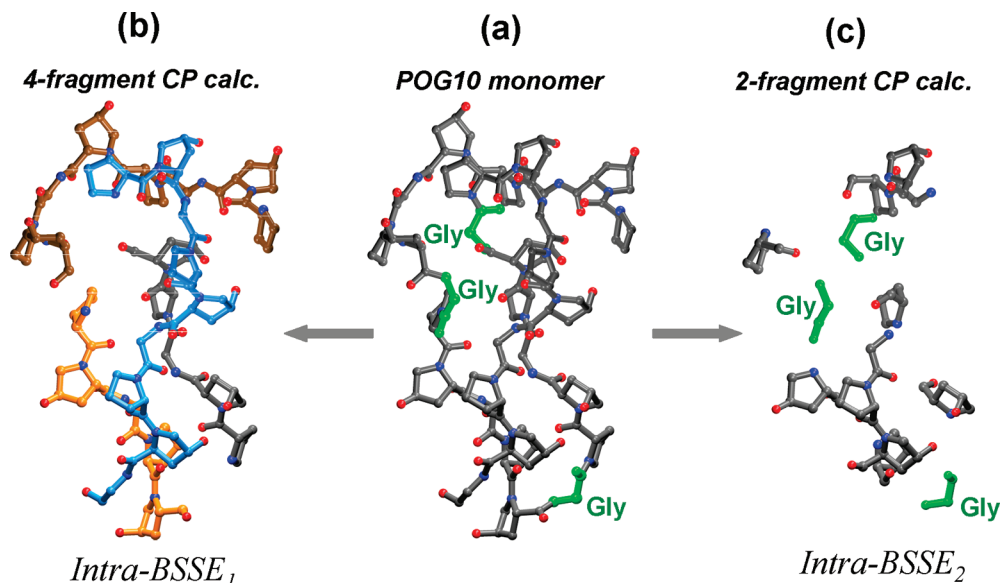


Figure 2. Ball-and-stick models of a **POG10** chain in its monomer state showing the fragmentation procedure followed to correct the *intramolecular* BSSE through CP calculations. (a) On the basis of a nonbonded interaction pair list, three glycine residues (in green) are selected in order to maximize the number of nonbonded interactions among the peptide fragments that result upon removal of the glycine residues. (b) BSSE arising from the interactions among the four peptide chains (C atoms are shown in different colors) is estimated using the CP procedure. (c) A molecular cluster is constructed from the coordinates of the glycine residues selected in a and those of the nearby peptide residues that interact directly with the marked glycines. The BSSE associated to the interaction between the glycines and the nearby groups is again estimated by means of CP calculations.

Table 3. Average Values and Standard Errors (in kcal/mol of peptide) of the Free-Energy Components for the Transition from the Monomeric to the Triple-Helix State at 300 K

	mean value	standard error		mean value	standard error
$\Delta \bar{E}_{\text{PBE/SVP}}$	53.7	9.7	$\Delta \bar{H}_{\text{MM-GBSA}}^{\text{norm}}$	0.7	0.1
$\Delta \bar{E}_{\text{PBE/SVP}}^{\text{CP-corrected}}$	49.8	9.6	$-T\Delta \bar{S}_{\text{MM-GBSA}}^{\text{norm}}$	8.8	0.9
$\Delta \bar{E}_{\text{PBE/SVP}}^{\text{elec}}$	-73.2	9.3	$-T\Delta \bar{S}_{\text{CP-corrected}}^{\text{conf}}$	0.4	-
$\Delta \Delta \bar{G}_{\text{COSMO}}^{\text{solute}}$	-10.7	0.4	$\Delta \bar{G}^{\text{CP-corrected}}$	-11.7	1.8
$\Delta \bar{E}_{\text{disp}}^{\text{solute-solvent}}$	11.0	0.8	$\Delta \bar{G}^a$	-7.8	2.1
$\Delta \bar{E}_{\text{disp}}^{\text{solute}}$	-1.2	0.1			
$\Delta \bar{G}_{\text{cav}}$					

^a Assuming a standard state of 0.001 M.

Free Energy for the Transition from Monomer to Triple Helix. As shown above, the fragment QM calculations complemented with the empirical dispersion formula can give insight into the nature of the interactions holding the peptide chains in the triple-helix conformation. However, the actual stability of the triple helix is determined by the free-energy change for dissociation to give the free peptide monomers. In our previous work,²⁹ we found that the isolated **POG10** peptide in aqueous solution adopts a stable folded conformation, and therefore, by combining the fragment QM data on the triple helix with the results of QM calculations on a representative set of **POG10** monomers, one could estimate the corresponding free-energy change for the peptide aggregation process leading to the **POG10** triple helix, provided that the selected QM method gives a compensated description of the conformational and intermolecular interaction energies. By taking advantage of our previous computational experience, we combined the QM energies with further molecular-mechanical data in order to ensure a balanced description of other free-energy components (solute-solvent vdW interactions, thermal contributions to free energy, etc.). More specifically, we used the following expression in order to

compute the average free energy of the **POG10** system both in its triple-helix and monomer states

$$\bar{G} = \bar{E}_{\text{PBE/SVP}}^{\text{CP-corrected}} + \bar{E}_{\text{disp}}^{\text{solute}} + \bar{E}_{\text{disp}}^{\text{solute-solvent}} + H_{\text{MM-GBSA}}^{\text{norm}} - T\bar{S}_{\text{MM-GBSA}}^{\text{norm}} + \Delta \bar{G}_{\text{COSMO}}^{\text{elec}} \quad (17)$$

where the gas-phase $\bar{E}_{\text{PBE/SVP}}^{\text{CP-corrected}}$ energy, which includes the intermolecular and intramolecular BSSE corrections, and the electrostatic solvation energy ($\Delta \bar{G}_{\text{COSMO}}^{\text{elec}}$) are computed by means of fragment-based (triple helix) and standard (monomer) QM calculations; the $\bar{E}_{\text{disp}}^{\text{solute}} + \bar{E}_{\text{disp}}^{\text{solute-solvent}}$ dispersion energy terms are computed with the same empirical formula, and normal mode molecular mechanical calculations are used to estimate the thermal contributions to free energy. The change in the average values of these energetic components for the monomer \rightarrow triple-helix transition are collected in Table 3, which also includes the corresponding small differences in the cavitation free energy and the conformational entropy that were computed following the procedures described in our previous work.²⁹

We see in Table 3 that the QM energy terms (gas-phase and solvation energy) as well as the empirical dispersion

energies change significantly on going from the monomer to the triple helix. In agreement with our previous molecular mechanical and Poisson–Boltzmann (MM-PB) calculations, the QM-based approach predicts also that the driving force for the formation of the triple helix is mainly provided by the electrostatic solvation energy. The total ΔG value obtained with the CP-corrected QM energies amounts to -11.7 kcal/mol, with a statistical uncertainty of 1.8 kcal/mol (standard error). This value is in moderate agreement with the most accurate experimental estimate at 300 K, -6.4 kcal/mol, which has been derived from differential scanning calorimetry.^{29,40} The purely MM-PB calculations together with a broader sampling give a ΔG value of -6.2 (1.2) kcal/mol.²⁹ The larger difference between the QM-based calculations and experiment is most likely due to several factors like the small error in the fragment-based QM calculations, the remaining inaccuracy in the correction of the intramolecular BSSE, slight unbalances in the combination of QM and MM data in eq 17, as well as by some limitations of the PBE DFT functional to reproduce the electrostatic and H-bond interactions. All these potential sources of error, which are not present in the MM-PB calculations, could be mitigated by gaining more computational experience and improving the details of the mixed QM–MM computational protocol. On the other hand, it turns out that the ΔG value obtained with the CP-uncorrected QM energies (-7.8 kcal/mol) is closer to the experimental estimate. Nevertheless, this result is somewhat fortuitous given that, in the particular case of the **POG10** system, the sum of intra- and intermolecular interactions remains approximately constant upon the monomer \rightarrow triple-helix transition.

Summary and Conclusions

In this work we reviewed several computational methods developed during the last years for computing the energy of large molecules using only fragment energies. Although some of the previous methods have been introduced independently to each other, a comparative analysis reveals their common roots, which, in our opinion, can be traced back to the general formalism of the MBE method. For biomolecules constructed with repetitive building blocks (residues, secondary structural elements,...), it is proposed that a simple thermochemical approach is probably the best option for formulating a *standard* fragment energy method. The validity of the fragment QM energy strategy has been tested intensively considering a challenging problem for simulation methodologies, that is, the prediction of the interchain interaction energy and the free energy for dissociation of a prototypical collagen model. The comparison of our fragment-based energies with experimental data and former theoretical results shows that the actual applicability of the fragment QM methods in biomolecular simulations will rely heavily on the proper combination of QM and MM calculations as well as in the conformational sampling performed by MM methods. Moreover, the correction of the inter- and intramolecular BSSE will be critically important for obtaining realistic energies of either interaction or conformational changes.

Since the MM-PB method predicts a more accurate value than the fragment-based QM calculations for the ΔG change

in the monomer \rightarrow triple-helix transition of the **POG10** system, one may raise the question of whether the fragment QM approaches are really needed. Clearly, the fragment QM calculations would have a broader applicability since they can be used to investigate all kinds of interactions and chemical transformations involving biomolecules. For example, most of the current force fields have been developed without specifically considering the interactions of biomolecules with metal ions, clusters, or surfaces, and therefore, the application of fragment-QM methodologies to study *biomaterials* could provide reliable energetic data, which in turn could be useful for the development and validation of new MM parameters. In addition, we point out that the QM charge densities obtained in the fragment calculations contain much valuable information that can be used for estimating other QM properties (e.g., electrostatic potential) and deriving QM descriptors (e.g., for determining ligand affinity). Similarly, the fragment QM calculations could also be used to outline electron pathways connecting the electron donor and acceptor sites in redox metalloproteins⁴¹ and the energy gaps between electronic states. Therefore, with the continuous improving in the efficiency of QM methodologies, the decreasing cost of computer hardware, as well as a necessary standardization of the fragment energy approach by means of intensive computational experimentation, the full QM description of large biomolecules could be done regularly in the near future.

Computational Methods

DFT Calculations. Density functional theory methods have become the most popular QM methodology for the study of biomolecules because they include electron correlation effects at a relatively cheap computational cost. In principle, the Perdew–Burke–Ernzerhoff (PBE)⁴² and Tao–Perdew–Staroverov–Scuseria (TPSS)⁴³ functionals are particularly attractive for performing fragment energy calculations, since they are nonempirical GGA functionals that give results with an acceptable quality in any type of chemical systems including macromolecules and condensed phases. In this work, we used the PBE functional combined with a double- ζ plus polarization basis set (SVP).⁴⁴ The reliability of the PBE/SVP level of theory was assessed by carrying out some validation calculations on a small triple-helix system (see below).

All DFT calculations were performed using the TURBOMOLE suite of programs,⁴⁵ in the framework of the multipole accelerated resolution-of-the-identity approximation (MARI-J) using the appropriate auxiliary basis set.^{46,47} To estimate the effect of the solvent environment on the DFT energies, we used the conductor-like screening model (COSMO) included in TURBOMOLE in which the solvent dielectric continuum is approximated by a scaled conductor.⁴⁸ The optimized atomic COSMO radii ($r_{\text{H}} = 1.3$ Å, $r_{\text{C}} = 2.0$ Å, $r_{\text{N}} = 1.83$ Å, and $r_{\text{O}} = 1.72$) were used to generate the solvent-accessible molecular cavity.⁴⁹ Note that in the thermochemical fragment energy calculations reported in this work long-range electrostatic effects are truncated in the different fragment calculations and that, therefore, a molecular cavity is constructed around each fragment system

Table 4. Average Values and Standard Deviations for the Interaction Energy (kcal/mol of THP) among the Three Peptide Chains for 25 Snapshots of the [Ace(Pro-Hyp-Gly)-Nme]₃ Trimer

level of theory	ΔE_{int}	level of theory	ΔE_{int}
PBE/SVP ^a	-10.7 ± 1.4	PBE/SVP ^b	-10.7 ± 6.2
PBE/TZVP ^a	-8.6 ± 1.3	PBE/TZVP ^b	-8.2 ± 5.8
PBE/TZVPP ^a	-9.0 ± 1.3		
TPSS/SVP ^a	-9.0 ± 1.1		

^a Geometries were extracted from the **POG10** MD simulations and relaxed via MM energy minimization. ^b Geometries were extracted from the **POG10** MD simulations.

$(Y_i - B_i^R, Y_{i-1} - A_i - B_i^R, \dots)$. This is fully consistent with the estimation of the full system energy from a combination of reaction energies (eqs R3 and R4).

Since the GGA density functionals are unable to describe dispersive interactions, the DFT energy terms were augmented with an dispersion energy contribution, E_{disp} , which was computed using an empirical formula that has been introduced by Elstner et al.³⁴ in order to extend their approximate DFT method for the description of dispersive interactions. The E_{disp} expression consists basically of a $-C_6/R^6$ term, which is appropriately damped for short R distances. We used the same parameters for C, N, O, and H and combination rules as those described by Elstner et al.⁴⁷

Molecular Geometries and Molecular Mechanical Calculations. Molecular geometries of the **POG10** system were taken from our previous study on the relative stability of collagen model peptides.²⁹ The triple-helix and monomer states of **POG10** were subject to 20 and 50 ns molecular dynamics (MD) simulations, respectively, at constant P (1 atm) and T (300 K) in explicit solvent using the AMBER package.⁵⁰ From these MD simulations, a set of 100 snapshots was extracted for each state and the internal geometry of the solute molecules was relaxed throughout energy minimization prior to the QM and MM energy calculations. The snapshots were postprocessed through the removal of all solvent molecules.

Thermal contributions to the enthalpy and entropy of solute molecules were estimated by means of MM normal mode calculations using the NAB package⁵¹ and following the prescriptions described elsewhere.²⁹ The nonpolar solvation energy was computed by combining the explicit solvent representation with an estimation of the relative change in the cavitation free energy of the solute.⁵² In our previous work, the conformational entropy of the solute was computed via an expansion of the so-called mutual information functions.⁶

Validation Calculations of the PBE/SVP Level of Theory. Table 4 summarizes the results of some preliminary validation calculations in which we computed the interchain interaction energy in a small THP model ([Ace-(Gly-Pro-Hyp)-Nme]₃; 123 atoms). In these calculations, we used the PBE and TPSS functionals combined with different basis sets ranging from the double- ζ SVP to the triple- ζ plus double polarization TZVPP. All DFT energies include the effect of aqueous solvent (COSMO model) and are combined with the empirical estimate of the dispersion energy. We also corrected the BSSE affecting the intermolecular interaction

energy by means of the counterpoise method. Coordinates of the small THP models were taken from 25 truncated snapshots of our previous MD simulations of the **POG10** system after having relaxed the internal geometry of the solute molecules via energy minimizations using the AMBER force field.

We see in Table 4 that the average PBE energies obtained with various basis sets are quite similar, the differences being around 1–2 kcal/mol. The TPSS functional gives similar interaction energies to those provided by PBE. By repeating some calculations without relaxing the internal geometry of the small THP models, we found that the average interaction energies are hardly affected, but standard deviations are much higher (~ 6 kcal/mol). Overall, we conclude that the PBE/SVP energy calculations on the MM-relaxed geometries may constitute a reasonable compromise between quality and computational cost.

Acknowledgment. This research was supported by the following grants: FICYT (Asturias, Spain) IB05-076 and MEC (Spain) CTQ2007-63266. E.S. and N.D. thank MEC for their FPU and Ramon y Cajal contracts, respectively. We are grateful to Dr. H. Valdés for her careful reading of the manuscript and suggestions.

Supporting Information Available: Equivalence between second-order MBE and KEM; Tables S1 and S2 containing the relative and absolute energies of all the MD structures considered in the test calculations. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Yang, W.; Lee, T.-S. *J. Chem. Phys.* **1995**, *103*, 5674.
- (2) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879.
- (3) Connolly, J. W. D.; Williams, A. R. *Phys. Rev. B* **1983**, *27*, 5169.
- (4) Carlsson, A. E. Beyond pair potentials in elemental transition metals and semiconductors. In *Solid State Physics*; Ehrenreich, H., Turnbull, D., Eds.; Academic Press: Boston, 1990; Vol. 43, p 1.
- (5) Drautz, R.; Fähnle, M.; Sanchez, J. M. *J. Phys.: Condens. Matter* **2004**, *16*, 3843.
- (6) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096.
- (7) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2005**, *103*, 808.
- (8) Huang, L.; Massa, L.; Karle, J. *Int. J. Quantum Chem.* **2006**, *106*, 447.
- (9) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2005**, *102*, 12690.
- (10) Huang, L.; Massa, L.; Karle, J. *J. Chem. Theory Comput.* **2007**, *3*, 1337.
- (11) Huang, L.; Massa, L.; Karle, J. *Proc. Nat. Acad. Sci. U.S.A.* **2008**, *105*, 1849.
- (12) Xantheas, S. S. *J. Chem. Phys.* **1994**, *100*, 7523.
- (13) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *20*, 6832.

- (14) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701.
- (15) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.
- (16) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 46.
- (17) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342.
- (18) Sorkin, A.; Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 683.
- (19) Fedorov, D. G.; K, K. *J. Phys. Chem. A* **2007**, *111*, 6904.
- (20) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.
- (21) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484.
- (22) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.
- (23) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215.
- (24) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112.
- (25) Collins, M. A.; Deevb, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.
- (26) Vreven, T.; Frisch, M. J.; Kudin, N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys.* **2006**, *104*, 701.
- (27) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760.
- (28) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889.
- (29) Suarez, E.; Diaz, N.; Suarez, D. *J. Phys. Chem. B* **2008**, *112*, 15248.
- (30) Brodsky, B.; Persikov, A. V. *Adv. Protein Chem.* **2005**, *70*, 301.
- (31) Bella, J.; Brodsky, B.; Berman, H. M. *Structure* **1995**, *3*, 893.
- (32) Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75.
- (33) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (34) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
- (35) Jureka, P.; Cerný, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.
- (36) Valdés, H.; Klusák, V.; Pitoák, M.; Exner, O.; Starý, I.; Hobza, P.; L., R. *J. Comput. Chem.* **2008**, *29*, 861.
- (37) Shields, A. E.; van Mourik, T. *J. Phys. Chem. A* **2007**, *111*, 13272.
- (38) Palermo, N. Y.; Csontos, J.; Owen, M. C.; Murphy, R. F.; Lovas, S. *J. Comput. Chem.* **2007**, *28*, 1208.
- (39) Asturiol, D.; Duran, M.; Salvador, P. *J. Chem. Phys.* **2008**, *128*, 144108.
- (40) Nishi, Y.; Uchiyama, S.; Doi, M.; Nishiuchi, Y.; Nakazawa, T.; Ohkubo, T.; Kobayashi, Y. *Biochemistry* **2005**, *44*, 6034.
- (41) Guallar, V. *J. Phys. Chem. B* **2008**, *112*, 13460.
- (42) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (43) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (44) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.
- (45) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.
- (46) Sierka, M.; Hogeckamp, A.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *118*, 9136.
- (47) Eichkorn, K.; Treutler, O.; Ohm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242*, 652.
- (48) Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187.
- (49) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. *J. Phys. Chem. A* **1998**, *102*, 5074.
- (50) Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.
- (51) Macke, T.; Case, D. A. Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids*; Leontes, N. B., SantaLucia, J. J., Eds.; American Chemical Society: Washington, DC, 1998; pp 379.
- (52) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238.

A Thermochemical Fragment Energy Method for
Biomolecules:
Application to a Collagen Model Peptide

Ernesto Suárez, Natalia Díaz, and Dimas Suárez()*

Departamento de Química Física y Analítica,
Universidad de Oviedo, 33006 Oviedo (Asturias), Spain.

dimas@uniovi.es

SUPPORTING INFORMATION

EQUIVALENCE BETWEEN 2nd-ORDER MBE AND KEM

Rearranging the indices in the double sum and summing over the three terms in the parentheses separately, the MBE equation for the pairwise approximation can be rewritten in the following form:

$$E_M = \sum_i^M E_i + \frac{1}{2} \sum_i^M \sum_{j \neq i}^M E_{ij} - \frac{1}{2} \sum_i^M \sum_{j \neq i}^M E_j - \frac{1}{2} \sum_i^M \sum_{j \neq i}^M E_i \quad [1]$$

In the last term of equation [1] there are $M-1$ possible j -values for each i -value. If we permute $i \leftrightarrow j$ in the third term, we obtain the same result, and we have:

$$E_M = \sum_i^M E_i + \frac{1}{2} \sum_i^M \sum_{j \neq i}^M E_{ij} - \frac{1}{2} (M-1) \sum_i^M E_i - \frac{1}{2} (M-1) \sum_i^M E_i \quad [2]$$

Taking $\sum_i^M E_i$ as a common factor, we can group the first, third and fourth terms in the right side

of the last expression:

$$E_M = \frac{1}{2} \sum_i^M \sum_{j \neq i}^M E_{ij} - (M-2) \sum_i^M E_i = \sum_{i=1}^{M-1} \sum_{j=i+1}^M E_{ij} - (M-2) \sum_{i=1}^M E_i \quad [3]$$

By running through the double sum in equation [3], we sum by rows the elements of the following upper diagonal matrix:

$$\begin{pmatrix} 0 & E_{12} & E_{13} & \dots & E_{1M} \\ 0 & 0 & E_{23} & \dots & E_{2M} \\ 0 & 0 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & E_{M-1,M} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad [4]$$

If we sum by the m -diagonals of this matrix, then the equation [3] becomes the KEM expression:

$$E_M = \sum_{m=1}^{M-1} \left(\sum_{i=1}^{M-m} E_{i,i+m} \right) - (M-2) \sum_{i=1}^M E_i \quad [5]$$

TableS1. Inter-chain interaction energies^(a) (ΔE_{int} , in kcal per mol of peptide) of all the THP fragments^(a) employed in the fragment energy calculations of [Ace-(Pro-Hyp-Gly)₄-Nme]₃. Values of the total interaction energies as obtained with the fragment-based (FRAG) and conventional (CONV) calculations are also indicated.

	$A_1 - B_1^R$	$Y_1 - A_2 - B_2^R$	$Y_2 - A_3 - B_3^R$	$Y_1 - B_1^R$	$Y_2 - B_2^R$	[Ace-(Pro-Hyp-Gly) ₄ -Nme] ₃	
						FRAG.	CONV.
1	-14.1	-13.8	-15.7	-7.0	-7.9	-28.7	-28.7
2	-14.5	-15.0	-14.1	-8.0	-7.5	-28.0	-28.0
3	-14.9	-15.3	-15.4	-7.7	-8.5	-29.4	-29.4
4	-14.2	-15.2	-14.7	-7.7	-7.6	-28.8	-28.8
5	-16.5	-15.9	-14.9	-8.5	-8.0	-30.8	-30.8
6	-14.5	-14.9	-15.2	-7.5	-7.8	-29.3	-29.4
7	-14.1	-14.3	-15.5	-6.9	-7.6	-29.4	-29.4
8	-14.7	-15.4	-15.1	-7.7	-8.0	-29.5	-29.5
9	-15.7	-15.6	-15.7	-7.7	-8.4	-30.8	-30.8
10	-14.7	-15.4	-15.0	-7.7	-7.9	-29.6	-29.6
11	-15.6	-15.4	-15.5	-8.1	-8.3	-30.2	-30.2
12	-14.5	-12.9	-13.6	-7.4	-6.6	-27.1	-27.1
13	-13.7	-14.6	-14.3	-7.4	-7.6	-27.6	-27.6
14	-14.3	-15.0	-14.8	-8.1	-7.2	-28.8	-28.8
15	-15.4	-16.1	-15.3	-8.5	-8.0	-30.4	-30.4
16	-14.9	-15.5	-15.5	-7.9	-8.3	-29.7	-29.7
17	-15.9	-15.5	-15.0	-8.4	-8.2	-29.9	-29.9
18	-15.5	-14.8	-14.9	-7.8	-7.7	-29.8	-29.8
19	-14.5	-15.5	-16.3	-7.5	-8.5	-30.2	-30.3
21	-15.2	-15.5	-15.6	-8.2	-7.9	-30.2	-30.2
22	-14.6	-14.7	-15.1	-7.6	-8.1	-28.7	-28.7
23	-15.6	-15.4	-15.1	-8.4	-7.6	-30.1	-30.1
24	-14.8	-16.2	-16.2	-8.8	-8.5	-29.9	-29.9
25	-15.0	-15.4	-15.3	-7.9	-8.1	-29.7	-29.8

(a) CP-corrected PBE/SVP COSMO energies combined with empirical E_{disp} .

TableS2. Absolute energies of all the THP fragments^(a) (E , in au) employed in the fragment energy calculations of [Ace-(Pro-Hyp-Gly)₄-Nme]₃.

Values of the total absolute energies as obtained with the fragment-based (FRAG) and conventional (CONV) calculations are also indicated.

	$A_1 - B_1^R$	$Y_1 - A_2 - B_2^R$	$Y_2 - A_3 - B_3^R$	$Y_1 - B_1^R$	$Y_2 - B_2^R$	[Ace-(Pro-Hyp-Gly) ₄ -Nme] ₃	
						FRAG.	CONV.
1	-6329.71538	-6329.71852	-6329.72314	-3536.89447	-3536.89879	-11915.36378	-11915.36390
2	-6329.71167	-6329.71633	-6329.72884	-3536.88542	-3536.89501	-11915.37641	-11915.37653
3	-6329.72948	-6329.73052	-6329.72082	-3536.90795	-3536.90057	-11915.37230	-11915.37242
4	-6329.74093	-6329.74197	-6329.74200	-3536.91185	-3536.90514	-11915.40791	-11915.40800
5	-6329.73182	-6329.72688	-6329.72716	-3536.89671	-3536.90106	-11915.38809	-11915.38825
6	-6329.72631	-6329.70847	-6329.70472	-3536.90093	-3536.87868	-11915.35989	-11915.36001
7	-6329.72204	-6329.72529	-6329.72364	-3536.89634	-3536.89792	-11915.37671	-11915.37686
8	-6329.72259	-6329.73674	-6329.73912	-3536.90278	-3536.90608	-11915.38959	-11915.38972
9	-6329.72666	-6329.73446	-6329.73005	-3536.90592	-3536.90144	-11915.38381	-11915.38393
10	-6329.72072	-6329.72771	-6329.70258	-3536.89760	-3536.90117	-11915.35224	-11915.35240
11	-6329.73131	-6329.73801	-6329.73436	-3536.90720	-3536.90592	-11915.39056	-11915.39068
12	-6329.73258	-6329.71861	-6329.72255	-3536.89849	-3536.89563	-11915.37962	-11915.37972
13	-6329.69943	-6329.72562	-6329.73205	-3536.89766	-3536.90171	-11915.35773	-11915.35781
14	-6329.71516	-6329.72657	-6329.73368	-3536.89528	-3536.90031	-11915.37982	-11915.37995
15	-6329.73292	-6329.73175	-6329.73381	-3536.90433	-3536.89824	-11915.39591	-11915.39602
16	-6329.73949	-6329.73861	-6329.73252	-3536.90713	-3536.90706	-11915.39643	-11915.39654
17	-6329.74023	-6329.73932	-6329.73430	-3536.91016	-3536.90451	-11915.39918	-11915.39933
18	-6329.71976	-6329.71689	-6329.71542	-3536.89799	-3536.89086	-11915.36322	-11915.36338
19	-6329.71235	-6329.70664	-6329.71826	-3536.88148	-3536.89500	-11915.36077	-11915.36090
21	-6329.73042	-6329.71739	-6329.72103	-3536.89970	-3536.89054	-11915.37860	-11915.37881
22	-6329.74411	-6329.72986	-6329.72906	-3536.90627	-3536.89872	-11915.39804	-11915.39821
23	-6329.72928	-6329.72291	-6329.71737	-3536.89718	-3536.89854	-11915.37384	-11915.37404
24	-6329.69112	-6329.71447	-6329.73682	-3536.88343	-3536.90550	-11915.35348	-11915.35364
25	-6329.72746	-6329.70632	-6329.70716	-3536.89789	-3536.88097	-11915.36208	-11915.36223

(a) CP-corrected PBE/SVP COSMO energies combined with empirical E_{disp} .

**2.2.1.2 Thermochemical Fragment Energy Method for Quantum Mechanical
Calculations on Biomolecules**

Ernesto Suárez, Natalia Díaz and Dimas Suárez

*Proceedings of the 2009 International Conference on Computational and Mathematical
Methods in Science and Engineering, Editor: J. Vigo-Aguiar. Gijón 2009. p1407-1416*

Thermochemical Fragment Energy Method for Quantum Mechanical Calculations on Biomolecules

Ernesto Suárez, Natalia Díaz and Dimas Suárez

Departamento de Química Física y Analítica. Universidad de Oviedo.

suarezernesto.uo@uniovi.es, diazfnatalia@uniovi.es,
dimas@uniovi.es,

Abstract

Herein, we present benchmarking & validation calculations of the recently proposed Thermochemical Fragment Energy Method. This method has been designed to compute the quantum mechanical energy of large biomolecular systems through a linear combination of subsystem (fragment) energies, which, in turn, can be computed using standard quantum chemical programs. A density functional level of theory combined with an implicit solvent model is employed to compute the total energy of a prototypical collagen model containing more than 1000 atoms. For this macromolecule, we find that our fragmentation scheme predicts reasonably accurate energies (the largest error being ~ 0.25 kcal/mol) and exhibits a large speedup (96%) over conventional calculations.

Keywords: Quantum Chemical Calculations, Linear Scaling Methods, Biomolecules

1. Introduction

During the last years, several methods have been developed in order to estimate the energy of a large molecule as a function of fragment energies. In this respect, the multi-body expansion (MBE) method is probably the most general way to evaluate the total energy as a *linear* combination of energies corresponding to isolated atomic clusters extracted from the global structure, in such a way that they include systematically two-, three-, and N -body effects.[1] In this formalism, the total energy of an M -particle system (composed of atoms, molecules, or

molecular fragments linked covalently) can be expressed as $E_M(A_1, A_2, \dots, A_M)$, where $A_i = \{\mathbf{R}_i, \sigma_i\}$ has the information about the coordinates (\mathbf{R}_i) and the type (σ_i) of the i -particle. By representing the total energy in terms of an expansion of a series of N -order (or N -body, or N -fragment) energy contributions, we have:

$$E_M(A_1, A_2, \dots, A_M) = \sum_{i=1}^M E(A_i) + \sum_{i<j}^M \Delta E^{(2)}(A_i, A_j) + \sum_{i<j<k}^M \Delta E^{(3)}(A_i, A_j, A_k) + \dots [1]$$

where $E(A_i)$ is the energy of the i -fragment, $\Delta E^{(2)}(A_i, A_j)$ is the interaction energy between the fragments i and j , and $\Delta E^{(3)}(A_i, A_j, A_k)$ is the additional energy due to three-body effects that cannot be assessed from a two body representation, and so on. Note that if eq. [1] is truncated up to second order, we reach the well-known pairwise additive approximation to the total energy.

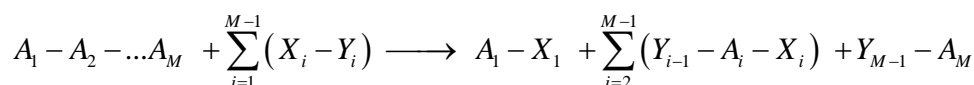
In general, a low-order MBE is not enough to accurately compute the total energy of complex systems and, unfortunately, the calculation of high-order MBE terms is extremely expensive in terms of computer time. Nevertheless, the MBE formalism constitutes the appropriate framework for developing different computational strategies aimed at the evaluation of the total energy of large systems from subsystem (fragment) energies. Thus, the so-called Kernel Energy Method, which has been utilized to compute the quantum mechanical (QM) energy of large biomolecules [2], is basically a direct truncation of the multi-body expansion. Similarly, the Fragment Molecular Orbital method is a more refined version of MBE in which the energies of the different fragments are computed by iteratively solving *effective* fragment Hamiltonians.[3] A cheaper alternative for non-covalently connected fragments is the Electrostatically Embedded Many-Body Expansion, [4] in which the energy of each cluster is calculated taking into account the electric field due to the fixed partial atomic charges of the surrounding fragments. The Molecular Fractionation with Conjugate Caps [5] and the Systematic Molecular Fragmentation[6] schemes have been justified by means of simple thermochemical arguments, but they can also be derived within the MBE formalism. The Molecular Tailoring Approach [7] is another example of a formal derivation within the MBE framework, in which the system is divided into overlapping fragments and, subsequently, the total energy is estimated by summing the fragment contributions and then subtracting the energies of fragment intersections.

Very recently, on the basis of thermochemical arguments, we have proposed yet another variant of the fragment-based computational methods, the so-called Thermochemical Fragment Energy Method (TFEM),[8] which can be particularly useful for biomolecules using either quantum mechanical (QM) or hybrid quantum mechanical/molecular mechanics (QM/MM) methods. In this work, we will first outline the essentials of the TFEM technique. Then, we will present a

series of benchmarking & validation calculations on a complex biomolecule in aqueous solution. Finally, we will discuss the potential applicability of the TFEM approach for computing accurate QM energies of large biomolecules.

2. Thermochemical Fragment Energy Method

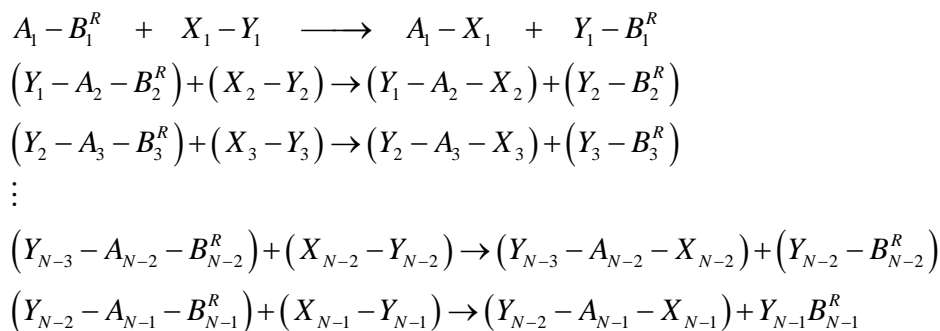
Let us consider the rupture of a macromolecule P composed of M fragments A_i ($P \equiv A_1 - A_2 - \dots - A_M$) through the following formal process:



where X_i and Y_i stand for all the molecular caps that are required to saturate the exposed bonds. The total energy change corresponding to the above formal reaction is:

$$\Delta E = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + E(Y_{M-1} - A_M) - \sum_{i=1}^{M-1} E(X_i - Y_i) - E(\mathbf{P}) \quad [2]$$

The same ΔE change can be estimated by performing a series of formal fragmentation steps in which the corresponding reactants are defined on the basis of some geometric and/or chemical-structure criteria (*e.g.*, a cut-off distance R) as shown in the following chemical equations:



where B_i^R represents a *buffer* region, which includes all the neighboring atoms (or fragments) that are around A_i depending on the R -criteria being used. The sum of the $M-1$ fragmentation processes defined in this manner leads to the following chemical equation:

$$A_1 - B_1^R + \sum_{i=2}^{M-1} (Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1} (X_i - Y_i) \longrightarrow$$

$$A_1 - X_1 + \sum_{i=2}^{M-1} (Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1} (Y_i - B_i^R)$$

In this way, the energy change for the total fragmentation of \mathbf{P} through the R -dependant fragmentation processes (ΔE^R) is given by:

$$\Delta E^R = E(A_1 - X_1) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - X_i) + \sum_{i=1}^{M-1} E(Y_i - B_i^R) -$$

$$- \left[E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + \sum_{i=1}^{M-1} E(X_i - Y_i) \right] \quad [3]$$

Extracting the exact fragmentation energy ΔE from equation [2] and defining $\delta E = \Delta E^R - \Delta E$, we can combine equations [2] and [3] to express the *exact* total energy of the system $E(\mathbf{P})$ in terms of the fragment energies and the δE difference:

$$E(\mathbf{P}) = \left[E(A_1 - B_1^R) + \sum_{i=2}^{M-1} E(Y_{i-1} - A_i - B_i^R) + E(Y_{M-1} - A_M) \right] -$$

$$- \left[\sum_{i=1}^{M-1} E(Y_i - B_i^R) \right] + \delta E(\mathbf{B}^R, \mathbf{Y}) \quad [4]$$

where the δE difference is expressed as a function of $\mathbf{B}^R = \{B_i^R\}$ and $\mathbf{Y} = \{Y_i\}$. If we neglect the δE term in eq. [4], the resulting linear combination of fragment energies defines the Thermochemical Fragment Energy Method, which is, by construction, a linear scaling method. In principle, its accuracy can be controlled by increasing the R -criterion in order to include larger portions of the remaining \mathbf{P} molecule in the B_i^R buffer regions. Alternatively, we could use QM/MM methodologies in order to calculate the reaction energies of the fragmentation steps so that the R -criterion would be applied to select the size of the QM region.

3. Computational Details

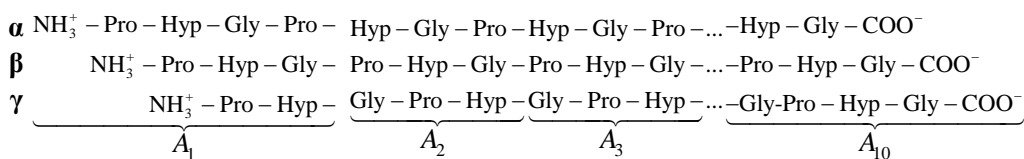
Given that density functional theory methods (DFT) have become the most popular QM methodology for the study of biomolecules, we used the Perdew-

Burke-Ernzerhoff (PBE) functional, which is a non-empirical functional that gives results with an acceptable quality in any type of chemical systems.[9] The PBE functional was combined with a double- ζ plus polarization basis set, which provides 14 Gaussian type functions (GTFs) centered on each heavy atom (2 GTFs per H atom).[10] All the DFT calculations were performed using the TURBOMOLE 5.9 suite of programs[11], in the framework of the multipole accelerated resolution-of-the-identity approximation (MARI-J) using the appropriate auxiliary basis set.[12] To estimate the effect of the solvent environment on the DFT energies, we used the conductor-like screening model (COSMO) included in TURBOMOLE in which the solvent dielectric continuum is approximated by a scaled conductor.[13]

All the calculations were run on the same hardware, HP BL465c servers equipped with two Quad Core Opteron 2356 processors and 32GB RAM memory, and with the same version of the software (Linux kernel 2.6.9). Due to the large size of the molecular systems (see below), the conventional DFT calculations on the whole molecules were carried out using the TURBOMOLE 5.9 binaries linked to the HP-MPI 2.3 library and running the calculations in parallel across 4 servers (32 cores) interconnected through an Infiniband fabric. On the other hand, the fragment-based energies were obtained by means of concurrent batches of serial TURBOMOLE executions on relatively small (fragment) subsystems.

4. Results

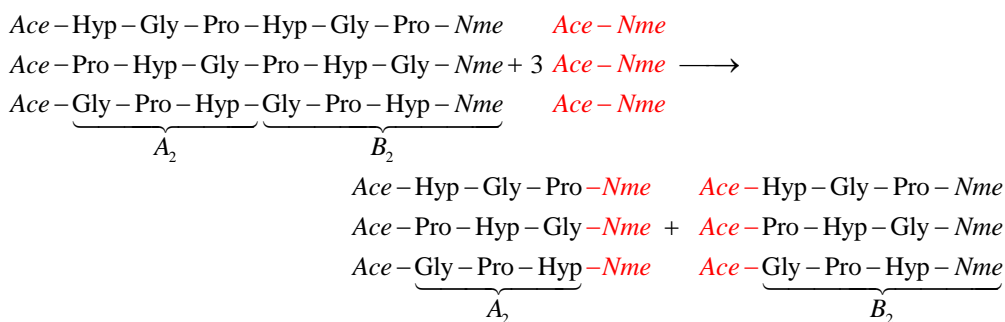
The test calculations reported in this work were performed on a prototypical collagen model, [(Pro-Hyp-Gly)₁₀]₃ (1089 atoms), which is a triple helical peptide (THP) that adopts an extended form in aqueous solution. This system is a trimer composed of three peptide chains (labeled as α , β and γ) with 30 amino acids per chain (see Scheme 1; Hyp is 4(*R*)-hydroxyproline). Molecular geometries were taken from our previous study on the relative stability of collagen model peptides.[14]



SCHEME 1

Although the fragment energy expression in eq. [4] is easily derived by assuming a linear macromolecule, it is also applicable for more complex macromolecules like the THPs. To this end, the triple helix can be better described as a linear arrangement of 10 fragments comprising each one three triplets of residues from the α , β and γ chains, respectively (see Scheme 1). On the other hand, Scheme 2 shows one example of the formal fragmentation processes required for the TFEM

calculations. Thus, the terminating groups attached to the *N*-terminal end of the A_2 block comprise three acetyl groups (*Ace*), whose coordinates are extracted from the previous A_1 fragment and augmented with the H-link atoms. Similarly, the corresponding B_3 buffer group includes the A_3 fragment plus three *N*-methyl (*Nme*) moieties extracted from the A_4 fragment. This formal fragmentation process can also be applied straightforwardly to obtain the energy of the individual peptide chains α , β , and γ .



SCHEME 2

The quality of the fragment energy calculations in order to reproduce the energetic properties of the full THP molecule can be assessed on the basis of data collected in Table 1. Note, however, that the conventional DFT calculations reported in this work are still far from being considered “routine” calculations owing to the large size of the THP molecule (1089 atoms, 9054 GTFs). For the THP system, the application of the thermochemical fragment partitioning protocol above described results in 17 fragment (subsystems) of varying size: the largest/smallest fragment having 252/144 atoms and 2052/1152 GTFs.

Interestingly, we see in Table 1 that the total energy of the whole THP system in aqueous solution can be computed accurately in terms of fragment energies given that the error in the fragment-based energies with respect to the exact full QM values is very small, only ~ 0.0001 au (~ 0.1 kcal/mol). It must also be noted that this error hardly depends either on the molecular geometry or the size of the molecular system, as very similar results are obtained in different geometries and in the α - β - γ peptide chains (see Table 1). Moreover, the inaccuracy in the TFEM energies is marginally superior to the numerical errors in the DFT energies arising from approximations made in the computation of molecular integrals.

The marginal errors committed by the TFEM approximation in the calculation of the absolute energies are not systematic. Thus, we found that the ΔE_{int} values, which are *relative* quantities, do not benefit from a partial cancellation of errors. On the contrary, the TFEM interaction energies have slightly larger errors of 0.10-0.25 kcal/mol, which are close to the sum of the errors corresponding to the

absolute energies involved in ΔE_{int} . Nevertheless, the total errors in the fragment-based ΔE_{int} energies are still very small, well below the 1.0 kcal/mol limit that is commonly considered the chemical accuracy limit.

The accuracy of the absolute TFEM energies for the THP systems (~ 0.0001 au) is significantly larger than that reported in previous fragment energy calculations (~ 0.001). However, most of these calculations have been carried out in the gas-phase whereas in the present TFEM calculations we also include electrostatic solvent effects by means of a continuum model (COSMO). This fact strongly suggests that solvent continuum methodologies can improve the accuracy of the fragment-based QM calculations because intramolecular long-range interactions are dampened out by the electrostatic screening exerted by the surrounding solvent continuum.

In terms of wall-clock time elapsed from start to end of the calculations, a single-point energy calculation on the THP systems using the fragment approach takes about 20 hours (1 core). The same energy value obtained with conventional QM calculations requires about 18 hours (32 cores running in parallel). This results in a noticeable 28x speedup in the performance of the TFEM calculations with respect to the conventional methods (the actual computer time savings in the different geometry/system varies only very slightly due to convergence issues; see Table 1). Moreover, besides computer time, the conventional calculations demand much larger resources in terms of memory, MPI-communications, hard disk space, etc. For example, a 42% reduction in the storage requirements of the QM wavefunction was observed when the TFEM approach is used, what can be quite remarkable because each conventional QM calculation on a THP molecule generates a 678 MB gzipped wavefunction file.

TABLE 1. Total energies in aqueous solution (in au) for the THP molecule and the interacting peptide chains, in five different molecular geometries (1-5), as obtained with conventional calculations and with the TFEM approximation. The inter-chain interaction energies (ΔE_{int} , in kcal/mol) are also shown. The unsigned difference between the conventional and TFEM energies in kcal/mol are given in parentheses. Wall times (in seconds) for the completion of the calculations are also indicated.

	CONVENTIONAL		TFEM (fragment-based)	
	Energy	Wall time (32 cores)	Energy (<i>error</i>)	Wall time (1 core)
Geometry 1				
THP	-28156.626294	67677	-28156.626275 (0.01)	73966
α -chain	-9385.431367	5669	-9385.431396 (0.02)	6123
β -chain	-9385.408847	5459	-9385.408908 (0.04)	5845
γ -chain	-9385.439342	6440	-9385.4394427 (0.05)	5954
ΔE_{int}	-217.58		-217.46 (0.12)	
Geometry 2				
THP	-28156.608411	60067	-28156.608471 (0.04)	72733
α -chain	-9385.432021	5943	-9385.432098 (0.04)	6163
β -chain	-9385.415841	6533	-9385.415898 (0.04)	5961
γ -chain	-9385.433738	6293	-9385.433818 (0.05)	6115
ΔE_{int}	-205.08		-204.98 (0.10)	
Geometry 3				
THP	-28156.592892	70565	-28156.592688 (0.13)	75067
α -chain	-9385.409960	5924	-9385.410025 (0.04)	6026
β -chain	-9385.432942	6433	-9385.432995 (0.03)	5860
γ -chain	-9385.428022	6217	-9385.428098 (0.05)	6117
ΔE_{int}	-202.04		-201.79(0.25)	
Geometry 4				
THP	-28156.613579	67677	-28156.613352 (0.14)	71623
α -chain	-9385.427485	5669	-9385.427527 (0.03)	6124
β -chain	-9385.397303	5459	-9385.397386 (0.05)	5831
γ -chain	-9385.432017	6440	-9385.432067 (0.03)	6058
ΔE_{int}	-223.88		-223.63 (0.25)	
Geometry 5				
THP	-28156.598276	74587	-28156.598184 (0.06)	70295
α -chain	-9385.420472	5943	-9385.420540 (0.04)	6187
β -chain	-9385.419252	6533	-9385.419283 (0.02)	5892
γ -chain	-9385.421285	6293	-9385.421341 (0.04)	6139
ΔE_{int}	-211.64		-211.48 (0.15)	

5. Summary and Conclusions

In this work we have performed a series of stringent test calculations that allow us to further assess the actual performance of the Thermochemical Fragment Energy Method in order to compute the total energy of biopolymers. This strategy, which can be considered as a thermochemical truncation of the classical multi-body expansion, approximates the total energy by combining successive fragmentation energies involved in the formal degradation of the biomolecule. All the required calculations have been done using a standard QM package and taking into account the effect of a solvent continuum.

From our calculations on the collagen model peptide, it turns out that the unsigned error introduced by the fragmentation technique into the absolute or relative QM energies is predicted to be very small (< 1 kcal/mol). We believe that this small error is a reasonable price to be paid for the large performance jump provided by the TFEM calculations with respect to conventional calculations on the whole molecule, which are still extremely costly in terms of computational resources. Hence we conclude that the TFEM approach constitutes a promising alternative for computing not only the QM energy of large biomolecules, but also their energy gradients and other molecular properties. However, we also note that real-case applications of TFEM will require most likely the combination of QM methodologies with empirical methods accounting for dispersion interactions as well as solving other technical issues like the intramolecular counterpoise correction of the Basis Set Superposition Error. Similarly, many computational experiments will have to be designed and performed before the full QM description of large biomolecules can be done regularly using standardized fragment-based methods.

6. References

- [1] R. DRAUTZ, M. FÄHNLE AND J. M. SANCHEZ, *General relations between many-body potentials and cluster expansions in multicomponent systems*, J. Phys.: Condens. Matter, **16** (2004), 3843-3852.
- [2] L. HUANG, L. MASSA AND J. KARLE, *The kernel energy method of quantum mechanical approximation carried to fourth-order terms*, Proc. Nat. Acad. Sci. U.S.A., **105** (2008), 1849-1854.
- [3] D. G. FEDOROV AND K. KITaura, *Extending the power of quantum chemistry to large systems with the fragment molecular orbital method*, J. Phys. Chem. A, **111** (2007), 6904-6914.

- [4] A. SORKIN, E. E. DAHLKE AND D. G. TRUHLAR, *Application of the electrostatically embedded many-body expansion to microsolvation of ammonia in water clusters*, J. Chem. Theory Comput., **4** (2008), 683-688.
- [5] S. LI, W. LI AND T. FANG, *An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules*, J. Am. Chem. Soc., **127** (2005), 7215-7226.
- [6] M. A. COLLINS AND V. A. DEEVB, *Accuracy and efficiency of electronic energies from systematic molecular fragmentation*, J. Chem. Phys., **125** (2006), 104104.
- [7] V. GANESH, R. K. DONGARE, P. BALANARAYAN AND S. R. GADRE, *Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies*, J. Chem. Phys., **125** (2006), 104109.
- [8] E. SUAREZ, N. DÍAZ AND D. SUÁREZ, *Thermochemical fragment energy method for biomolecules: application to a collagen model peptide*, J. Chem. Theory Comput., DOI: 10.1021/ct8005002 (2009).
- [9] J. P. PERDEW, K. BURKE AND M. ERNZERHOF, *Generalized gradient approximation made simple*, Phys. Rev. Lett., **77** (1996), 3865-3868.
- [10] A. SCHÄFER, H. HORN AND R. AHLRICH, *Fully optimized contracted gaussian basis sets for atoms Li to Kr.*, J. Chem. Phys., **97** (1992), 2571-2577.
- [11] R. AHLRICH, M. BÄR, M. HÄSER, H. HORN AND C. KÖLMEL, *Electronic structure calculations on workstation computers: The program system TURBOMOLE*, Chem. Phys. Lett., **162** (1989), 165-169.
- [12] M. SIERKA, A. HOGEKAMP AND R. AHLRICH, *Fast evaluation of the Coulomb potential for electron densities using multipole accelerated resolution of identity approximation.*, J. Chem. Phys., **118** (2003), 9136-9148.
- [13] A. SCHÄFER, A. KLAMT, D. SATTEL, J. C. W. LOHRENZ AND F. ECKERT, *COSMO Implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems*, Phys. Chem. Chem. Phys., **2** (2000), 2187-2193.
- [14] E. SUÁREZ, N. DÍAZ AND D. SUÁREZ, *Entropic control of the relative stability of triple-helical collagen peptide models*, J. Phys. Chem. B, **112** (2008), 15248-15255.

Conclusiones

En primer lugar, en el presente trabajo se han alcanzado las siguientes conclusiones metodológicas:

1. Se ha mostrado que el cálculo de la entropía total de una molécula a temperatura ambiente (excluyendo translación y rotación) como suma de entropía vibracional media y conformacional pura, esta última obtenida a partir de simulaciones de dinámica molecular, es un protocolo robusto y aplicable a distintos sistemas moleculares. Su aplicación para alcanos en fase gas reproduce satisfactoriamente las respectivas entropías experimentales.
2. Habiendo analizado formalmente y comprobado en la práctica las limitaciones y el elevado coste del método MIE, en la evaluación de la entropía conformacional de sistemas moleculares de tamaño medio, se han diseñado y validado una serie de métodos teóricos relacionados, que superan en gran medida los inconvenientes del método MIE original.
 - 2.1. En un primer avance metodológico para implementar de una manera eficiente la expresión MIE en sistemas de cálculo, se reformuló dicha expansión eliminando su redundancia, de modo que la entropía de cada subsistema considerado sólo se calcula una vez. No obstante, esta reformulación es insuficiente para tratar sistemas moleculares como los modelos de colágeno.
 - 2.2. Aunque la combinación de la expansión MIE con un criterio de *cutoff* es casi trivial, no se ha encontrado un procedimiento exacto para eliminar su redundancia. Alternativamente, se ha propuesto el método AMIE con el que se logra una ganancia en eficiencia de hasta dos órdenes de magnitud. Como desventaja, las entropías AMIE para un *cutoff* dado no son invariantes ante la permutación en los índices de dos torsiones cualesquiera. Sin embargo, se ha encontrado que los errores cometidos son numéricamente muy pequeños y generalmente asumibles utilizando un protocolo adecuado.

- 2.3.** Partiendo del método AMIE, se ha derivado rigurosamente el método MLA, que da resultados casi idénticos a los de la expansión original MIE a todos los órdenes compatibles con el *cutoff* dado. El método MLA conserva la generalidad de la expansión MIE, pero elimina sus principales inconvenientes (véase conclusión 5).
- 2.4.** En el contexto de los cálculos de entropía conformacional a partir de simulaciones moleculares, se ha construido un estimador que filtra la falsa correlación (CC-MLA) y se ha establecido un criterio para seleccionar un *cutoff* óptimo para cada trayectoria y sistema particular analizado.
- 3.** Se ha desarrollado un programa llamado CENCALC que permite calcular eficientemente la entropía conformacional pura de una molécula a partir de simulaciones moleculares. En este programa están implementados todos los métodos para el cálculo de entropía conformacional desarrollados durante la Tesis.
- 4.** Se ha propuesto un método denominado TFEM para estimar las energías mecanocuánticas de biomoléculas a partir de la combinación lineal de las energías de sus fragmentos. Debido a que es un método que se construye a través del planteamiento formal de reacciones químicas de fragmentación, nos permite incluir fácilmente el efecto del disolvente empleando modelos de disolvente continuo, hecho que incrementa sensiblemente la exactitud de las energías. Por su naturaleza, este protocolo es fácilmente paralelizable y sería igualmente aplicable con métodos híbridos QM/MM, tanto para obtener la energía como sus gradientes y segundas derivadas con respecto a las posiciones nucleares.

Mediante la aplicación de distintas técnicas de simulación de biomoléculas de la Química Computacional, junto con la aplicación de la metodología desarrollada en esta Tesis, se han alcanzado las siguientes conclusiones sobre la estructura y estabilidad de los modelos de colágeno:

5. La triple hélice de los modelos de colágeno es una estructura rígida a nivel local, en cuanto a los cambios conformacionales en los diedros del *backbone*, pero que presenta una cierta flexibilidad global asociada principalmente a los movimientos de *bending* (flexión) de toda la molécula. Las hebras aisladas, por su parte, son estructuras mucho más flexibles, y dicha flexibilidad está directamente relacionada con el contenido en residuos Prolina e Hidroxiprolina.
6. El campo de fuerzas AMBER03 reproduce razonablemente bien las propiedades estructurales de los modelos de colágeno POG10 y T3-785, tanto a nivel local como global. Por otra parte, las simulaciones indican que los modelos sintéticos de colágeno THP-1 y fTHP-5 presentan los mismos patrones geométricos característicos observados para POG10 y T3-785. En los cuatro modelos se observa que su grado de helicidad depende claramente de la secuencia de residuos.
7. La formación de la triple hélice en modelos de colágeno conlleva una importante reducción en la entropía conformacional debido a la pérdida de flexibilidad de las tres cadenas peptídicas. Esta penalización es especialmente importante en modelos con una baja presencia de imido-ácidos en su secuencia. Análogamente, el cambio de entropía conformacional de substratos peptídicos en su unión a enzimas MMPs es importante, y su estimación puede dar lugar a predicciones más fiables sobre su afinidad relativa.
8. Se ha visto en el caso de la triple hélice prototípica POG10, que es posible predecir la energía mecanocuántica en disolución acuosa (y por tanto también otras propiedades) de modelos colágeno con cientos de átomos, aplicando el protocolo TFEM sobre fragmentos apropiadamente seleccionados. El aprovechamiento efectivo de los cálculos TFEM pasa por su combinación con otros términos energéticos y/o entrópicos, así como por la corrección del error por superposición de bases. Aunque la forma cuasi-lineal de los modelos de colágeno es particularmente favorable a la aplicación de métodos basados en fragmentos, los buenos resultados obtenidos sugieren que mediante una experimentación computacional sistemática podrían diseñarse protocolos tipo TFEM aplicables a todo tipo de biomoléculas.

Informe sobre el Factor de Impacto de las Publicaciones Presentadas

En virtud de la actual normativa para estudios de tercer ciclo de la Universidad de Oviedo, las Tesis depositadas bajo la modalidad de compendio de publicaciones deben incluir un informe con el factor de impacto de las publicaciones presentadas. En nuestro caso particular, las publicaciones aceptadas o enviadas se recogen en cinco revistas científicas indexadas en la base de datos *Journal Citation Reports* del ISI e incluidas en la categoría de revistas de Química Física o de otras áreas de conocimiento afines (véase la siguiente Tabla). Debe destacarse que los principales resultados y desarrollos figuran en revistas situadas en el primer cuartil de las respectivas categorías, como por ejemplo el *Journal of Chemical Theory and Computation*.

Revista	Año	Factor de Impacto	Categoría	Posición en el Ranking	Publicac. de la Tesis
<i>J. Chem. Theory Comput.</i>	2011	–		–	2+1 ^(enviado)
	2010	5.138	CHEMISTRY, PHYSICAL	18/127	–
	2009	4.804		18/121	1
<i>J. Phys. Chem. B</i>	2008	4.189	CHEMISTRY, PHYSICAL	22/113	1
<i>J. Chem. Inform. Model.</i>	2011	–	COMP. SCIEN., INTERDISCIP.	–	1 ^(enviado)
	2010	3.822		2/97	–
<i>Proteins</i>	2010	2.813	BIOCHEM. AND MOL. BIOL.	142/160	1
<i>Entropy</i>	2010	1.109	PHYSICS, MULTIDISCIP.	34/80	1

Tabla5.1: Factores de impacto y ranking de las revistas científicas en las que se han publicado los principales resultados de la presente Tesis. (Fuente: *Science Citation Report*; *ISI Web of Knowledge*)

Bibliografía

1. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., and Karplus, M. (1983) Charmm: A Program for Macromolecular Energy, Minimization, and Molecular Dynamics Calculations, *J. Comput. Chem.* *4*, 187-217.
2. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1984) Molecular Dynamics with Coupling to an External Bath, *J. Chem. Phys.* *81*, 3684-3690.
3. Karplus, M., Ichiye, T., and Pettit, B. M. (1987) Configurational Entropy of Native Proteins, *Biophys. J.* *52*, 1083-1085.
4. Jorgensen, W. L., and Tirado-Rives, J. (1988) The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin, *J. Am. Chem. Soc.* *110*, 1657-1666.
5. Sharp, K., and Honig, B. (1991) Electrostatic Interactions in Macromolecules: Theory and Applications, *Ann. Rev. Biophys. Biophys. Chem.* *19*, 301-332.
6. Olsson, M. H. M., Parson, W. W., and Warshel, A. (2006) Dynamical Contributions to Enzyme Catalysis: Critical Tests of A Popular Hypothesis, *Chem. Rev.* *106*, 1737-1756.
7. Oostenbrink, C., and Gunsteren, W. F. v. (2005) Free Energies of Ligand Binding for Structurally Diverse Compounds, *Proc. Natl. Acac. Sci. USA* *102*, 6750-6754.
8. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wrighers, W. (2010) Atomic-Level Characterization of the Structural Dynamics of Proteins, *Science* *330*, 341-346.
9. Karplus, M., and Kuriyan, J. (2005) Molecular Dynamics and Protein Function, *Proc. Natl. Acac. Sci. USA* *102*, 6679-6685.
10. Adcock, S. A., and McCammon, J. A. (2006) Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins, *Chem. Rev.* *106*, 1589-1615.
11. Daggett, V. (2006) Protein Folding-Simulation, *Chem. Rev.* *106*, 1898-1916.
12. Andricioaei, I., and Karplus, M. (2001) On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations, *J. Chem. Phys.* *115*, 6289-6292.
13. Baron, R., Gunsteren, W. F. v., and Hünenberger, P. H. (2006) Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation, *Trends in Physical Chemistry* *11*, 88-122.
14. Grünberg, R., Nilges, M., and Leckner, J. (2006) Flexibility and Conformational Entropy in Protein-Protein Binding, *Structure* *14*, 683-693.
15. Killian, B. J., Yudenfreund-Kravitz, J., Somani, S., Dasgupta, P., Pang, Y.-P., and Gilson, M. K. (2009) Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an HIV-Derived PTAP Nonapeptide, *J. Mol. Biol.* *389*, 315-335.

16. Meirovitch, H. (2009) Absolute Free Energy and Entropy of a Mobile Loop of the Enzyme Acetylcholinesterase, *J. Phys. Chem. B* *113*, 7950-7964.
17. Zhou, H.-X., and Gilson, M. K. (2009) Theory of Free Energy and Entropy in Noncovalent Binding, *Chem. Rev.* *109*, 4092-4107.
18. Freddolino, P. L., Harrison, C. B., Liu, Y., and Schulten, K. (2010) Challenges in Protein Folding Simulations, *Nat. Phys.* *6*, 751-758.
19. Woo, H.-J., and Roux, B. (2005) Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations, *Proc. Natl. Acac. Sci. USA* *102*, 6825–6830.
20. Field, M. J. (2002) Simulating Enzyme Reactions: Challenges and Perspectives, *J. Comput. Chem.* *23*, 48-58.
21. Kuhn, B., and Kollman, P. A. (2000) Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models, *J. Med. Chem.* *43*, 3786-3791.
22. Gohlke, H., Kiel, C., and Case, D. A. (2003) Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes, *J. Mol. Biol.* *330*, 891-913.
23. Gohlke, H., and Case, D. A. (2003) Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein–Protein Complex Ras–Raf, *J. Comput. Chem.* *25*, 238-250.
24. Fitter, J. (2003) A Measure of Conformational Entropy Change during Thermal Protein Unfolding Using Neutron Spectroscopy, *Biophys. J.* *84*, 3924-3930.
25. Bachmann, A., Kiefhaber, T., Boudko, S., Engel, J., and Bächinger, H. P. (2005) Collagen Triple-Helix Formation in All-Trans Chains Proceeds by a Nucleation Growth Mechanism with a Purely Entropic Barrier, *Proc. Natl. Acac. Sci. USA* *102*, 13897-13902.
26. Suárez, E., Díaz, N., and Suárez, D. (2008) Entropic Control of the Relative Stability of Triple-helical Collagen Peptide Models, *J. Phys. Chem. B* *112*, 15248–15255.
27. Stone, M. J. (2001) NMR Relaxation Studies of the Role of Conformational Entropy in Protein Stability and Ligand Binding, *Acc. Chem. Res.* *34*, 379-388.
28. Villà, J., Strajbl, M., Glennon, T. M., Sham, Y. Y., Chu, Z. T., and Warshel, A. (2000) How Important are Entropic Contributions to Enzyme Catalysis?, *Proc. Natl. Acac. Sci. USA* *97*, 11899-11904.
29. Karplus, M., and Kushick, J. N. (1981) Method for Estimating the Configurational Entropy of Macromolecules, *Macromolecules* *14*, 325-332.
30. Schlitter, J. (1993) Estimation of Absolute and Relative Entropies of Macromolecules using the Covariance Matrix, *Chem. Phys. Lett.* *215*, 617–621.
31. Hnizdo, V., Darian, E., Fedorowicz, A., Demchuk, E., Li, S., and Singh, H. (2006) Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules, *J. Comput. Chem.* *28*, 655–668.
32. Killian, B. J., Kravitz, J. Y., and Gilson, M. K. (2007) Extraction of configurational entropy from molecular simulations via an expansion approximation, *J. Chem. Phys.* *127*.
33. Li, D.-W., and Brüschweiler, R. (2009) In silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides, *Phys. Rev. Lett.* *102*, 118108.

34. Hensen, U., Lange, O. F., and Grubmüller, H. (2010) Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach, *PLoS ONE* 5, e9179.
35. Kirkwood, J. G. (1935) Statistical Mechanics of Fluid Mixtures, *J. Chem. Phys.* 3, 300-313.
36. Kirkwood, J. G. (1968) *Theory of Liquids*, Gordon and Breach, New York.
37. Numata, J., Wan, M., and Knapp, E. (2007) Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation, *Genome Inform.* 18, 192-205.
38. Baron, R., Hünenberger, P. H., and McCammon, J. A. (2009) Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties, *J. Chem. Theory Comput.* 5, 3150–3160.
39. Hnizdo, V., Tan, J., Killian, B. J., and Gilson, M. K. (2008) Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods, *J. Comput. Chem.* 29, 1605–1614.
40. Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003) Nearest Neighbor Estimates of Entropy, *Am. J. of Math. Manag. Sci.* 23, 301-321.
41. Hnizdo, V., Darian, E., Fedorowicz, A., Demchuk, E., Li, S., and Singh, H. (2007) Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules, *J. Comput. Chem.* 28, 655–668.
42. Hensen, U., Grubmüller, H., and Lange, O. F. (2009) Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-dimensional Configuration Spaces, *Phys. Rev. E* 80, 011913.
43. Tyka, M., Clarke, A., and Sessions, R. (2006) An Efficient, Path-Independent Method for Free-Energy Calculations, *J. Phys. Chem. B* 110, 17212-17220.
44. Killian, B. J., Kravitz, J. Y., and Gilson, M. K. (2007) Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation, *J. Chem. Phys.* 127, 024107.
45. Schäfer, H., Mark, A. E., and Gunsteren, W. F. v. (2000) Absolute Entropies from Molecular Dynamics Simulation Trajectories, *J. Chem. Phys.* 113, 7809-7817.
46. Chang, C., Chen, W., and Gilson, M. K. (2005) Evaluating the Accuracy of the Quasiharmonic Approximation, *J. Chem. Theory Comput.* 1, 1017.
47. Ben-Naim, A. (2008) *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific, Singapore.
48. Matsuda, H. (2000) Physical Nature of Higher-Order Mutual Information: Intrinsic Correlations and Frustration, *Phys. Rev. E* 62, 3098-3102.
49. Cheluvvaraja, S., and Meirovitch, H. (2004) Simulation Method for Calculating the Entropy and Free Energy of Peptides and Proteins, *Proc. Natl. Acac. Sci. USA* 101, 9241-9246.
50. White, R. P., and Meirovitch, H. (2004) A Simulation Method for Calculating the Absolute Entropy and Free Energy of Fluids: Application to Liquid Argon and Water, *Proc. Natl. Acac. Sci. USA* 101, 9235-9240.
51. Chang, C. A., Chen, C., and Gilson, M. K. (2007) Ligand Configurational Entropy and Protein Binding, *Proc. Natl. Acac. Sci. USA* 104, 1534–1539.
52. Carlsson, A. E. (1990) *Solid State Physics*, Vol. 43, Boston.

53. Drautz, R., Fähnle, M., and Sanchez, J. M. (2004) General Relations Between Many-Body Potentials and Cluster Expansions in Multicomponent Systems, *J. Phys. Condens. Matter* 16, 3843.
54. Comtet, L. (1974) *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, Reidel Publishing Company, Holland.
55. Rohatgi, V. K. (2003) *Statistical Inference*, Dover, New York.
56. Huang, L., Massa, L., and Karle, J. (2005) Kernel Energy Method Illustrated With Peptides, *Int. J. Quantum Chem.* 103, 808.
57. Huang, L., Massa, L., and Karle, J. (2006) Kernel Energy Method: Basis Functions and Quantum Methods, *Int. J. Quantum Chem.* 106, 447-457.
58. Huang, L., Massa, L., and Karle, J. (2005) Kernel Energy Method: Application to Insulin, *Proc. Natl. Acad. Sci. USA* 102, 12690–12693.
59. Huang, L., Massa, L., and Karle, J. (2007) Kernel Energy Method: The Interaction Energy of the Collagen Triple Helix, *J. Chem. Theory Comput.* 3, 1337-1341.
60. Huang, L., Massa, L., and Karle, J. (2008) The Kernel Energy Method of Quantum Mechanical Approximation Carried to Fourth-Order Terms, *Proc. Natl. Acad. Sci. USA* 105, 1849–1854.
61. Suárez, E., Díaz, N., and Suárez, D. (2009) Thermochemical Fragment Energy Method for Biomolecules: Application to a Collagen Model Peptide, *J. Chem. Theory Comput.* 5, 1667–1679.
62. Fedorov, D. G., and Kitaura, K. (2004) The Importance of Three-Body Terms in the Fragment Molecular Orbital Method, *J. Chem. Phys.* 20, 6832.
63. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., and Uebayasi, M. (1999) Fragment Molecular Orbital Method: An Approximate Computational Method for Large Molecules, *Chem. Phys. Lett.* 313, 701-706.
64. Kitaura, K., Sugiki, S.-I., Nakano, T., Komeiji, Y., and Uebayasi, M. (2001) Fragment Molecular Orbital Method: Analytical Energy Gradients. , *Chem. Phys. Lett.* 336, 163-170.
65. Dahlke, E. E., and Truhlar, D. G. (2007) Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters *J. Chem. Theory Comput.* 3, 46 -53.
66. Dahlke, E. E., and Truhlar, D. G. (2007) Electrostatically Embedded Many-Body Correlation Energy, with Applications to the Calculation of Accurate Second-Order Miller-Plesset Perturbation Theory Energies for Large Water Clusters *J. Chem. Theory Comput.* 3, 1342 -1348.
67. Sorkin, A., Dahlke, E. E., and Truhlar, D. G. (2008) Application of the Electrostatically Embedded Many-Body Expansion to Microsolvation of Ammonia in Water Clusters, *J. Chem. Theory Comput.* 4 683–688.
68. Zhang, D. W., and Zhanga, J. Z. H. (2003) Molecular Fractionation with Conjugate Caps for Full Quantum Mechanical Calculation of Protein–Molecule Interaction Energy, *J. Chem. Phys.* 119, 3599.
69. Li, S., Li, W., and Fang, T. (2005) An Efficient Fragment-Based Approach for Predicting the Ground-State Energies and Structures of Large Molecules, *J. Am. Chem. Soc.* 127, 7215-7226.
70. Jiang, N., Ma, J., and Jiang, Y. (2006) Electrostatic Field-Adapted Molecular Fractionation with Conjugated Caps for Energy Calculations of Charged Biomolecules, *J. Chem. Phys.* 124, 114112.

71. Ganesh, V., Dongare, R. K., Balanarayan, P., and Gadre, S. R. (2006) Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies, *J. Chem. Phys.* *125*, 104109.
72. Collins, M. A., and Deevb, V. A. (2006) Accuracy and efficiency of electronic energies from systematic molecular fragmentation, *J. Chem. Phys.* *125*, 2006, 104104.
73. Mayhall, N. J., and Raghavachari, K. (2011) Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials, *J. Chem. Theory Comput.* *7*, 1336-1343.
74. Svensson, M., Humbel, S., Froese, R. D. J., Matsubara, T., Sieber, S., and Morokuma, K. (1996) ONIOM: A Multilayered Integrated MO+MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels–Alder Reactions and Pt(P(t-Bu)₃)₂ + H₂ Oxidative Addition, *J. Phys. Chem.* *100*, 19357-19363.
75. Suárez, E., Suárez, D., and Díaz, N. (2009) Thermochemical Fragment Energy Method for Quantum Mechanical Calculations on Biomolecules, In *Proceedings of the 2009 International Conference on Computational and Mathematical Methods in Science and Engineering* (Vigo-Aguiar, J., Ed.), pp 1407-1416, Gijón.
76. Schor, S. L. (1980) Cell Proliferation and Migration on Collagen Substrata in Vitro, *J. Cell Sci.* *41*, 159-175.
77. Streuli, C. (1999) Extracellular Matrix Remodelling and Cellular Differentiation, *Curr. Opin. Cell Biol.* *11*, 634-640.
78. Brodsky, B., and Shah, N. K. (1995) The Triple-Helix Motif in Proteins, *The FASEB Journal* *9*, 1537.
79. Rich, A., and Crick, F. H. C. (1961) The Molecular Structure of Collagen, *J. Mol. Biol.* *3*, 483-506.
80. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and Molecular Structure of a Collagen-like Peptide at 1.9Å Resolution, *Science* *266*, 75-81.
81. Kramer, R. Z., Bella, J., Brodsky, B., and Berman, H. M. (2001) The Crystal and Molecular Structure of a Collagen-like Peptide With A Biologically Relevant Sequence, *J. Mol. Biol.* *311*, 131-147.
82. Traore, A., Foucat, L., and Renou, J. P. (2000) ¹H-NMR Study of Water Dynamics in Hydrated Collagen: Transverse Relaxation-Time and Diffusion Analysis, *Biopolymers* *53*, 476-483.
83. Melacini, G., Bonvin, A. M. J. J., Goodman, M., Boelens, R., and Kaptein, R. (2000) Hydration Dynamics of the Collagen Triple Helix by NMR, *J. Mol. Biol.* *300*, 1041-1048.
84. Go, N., and Suezaki, Y. (1972) Analysis of the Helix-Coil Transition in (Pro-Pro-Gly)_n by the All-or-None Model, *Biopolymers* *12*, 1927-1930.
85. Brodsky, B., Thiagarajan, G., Madhan, B., and Kar, K. (2008) Triple-Helical Peptides: An Approach to Collagen Conformation, Stability, and Self-Associaton, *Biopolymers* *89*, 345.
86. Ottl, J., and Moroder, L. (1999) Disulfide-Bridged Heterotrimeric Collagen Peptides Containing the Collagenase Cleavage Site of Collagen Type I. Synthesis and Conformational Properties, *J. Am. Chem. Soc.* *121*, 653-661.
87. Nishi, Y., Uchiyama, S., Doi, M., Nishiuchi, Y., Nakazawa, T., Ohkubo, T., and Kobayashi, Y. (2005) Different Effects of 4-Hydroxyproline and 4-

- Fluoroproline on the Stability of Collagen Triple Helix, *Biochemistry* 44, 6034-6042.
88. Beck, K., Chan, V. C., Shenoy, N., Kirkpatrick, A., Ramshaw, J. A. M., and Brodsky, B. (1999) Destabilization of Osteogenesis Imperfecta Collagen-Like Model Peptides Correlates with the Identity of the Residue Replacing Glycine, *Proc. Natl. Acad. Sci. USA* 97, 4273-4278.
 89. Okuyama, K., Hongo, C., Fukushima, R., Wu, G., Narita, H., Noguchi, K., Tanaka, Y., and Nishino, N. (2004) Crystal Structures of Collagen Model Peptides with Pro-Hyp-Gly Repeating Sequence at 1.26 Å Resolution: Implications for Proline Ring Puckering, *Biopolymers* 76, 367-377.
 90. Li, M.-H., Fan, P., Brodsky, B., and Baum, J. (1994) Two-Dimensional NMR Assignments and Conformation of (Pro-Hyp-Gly)₁₀ and a Designed Collagen Triple-Helical Peptide, *Biochemistry* 32, 7377-7387.
 91. Berisio, R., Vitagliano, L., Mazzarella, L., and Zagari, A. (2001) Crystal Structure of a Collagen-Like Polypeptide with Repeating Sequence Pro-Hyp-Gly at 1.4 Å Resolution: Implications for Collagen Hydration, *Biopolymers* 56, 8-13.
 92. Lauer-Fields, J. L., Tuzinski, K. A., Shimokawa, K.-i., Nagase, H., and Fields, G. B. (2000) Hydrolysis of Triple-helical Collagen Peptide Models by Matrix Metalloproteinases, *J. Biol. Chem.* 275, 13282-13290.
 93. Lauer-Fields, J. L., Kele, P., Sui, G., Nagase, H., Leblanc, R. M., and Fields, G. B. (2003) Analysis of matrix metalloproteinase triple-helical peptidase activity with substrates incorporating fluorogenic l- or d-amino acids, *J. Biol. Chem.* 278, 105-115.
 94. Engel, J., Chen, H.-T., Prockop, D. J., and Klump, H. (1977) The Triple Helix-Coil Conversion of Collagen-Like Polytripeptides in Aqueous and Nonaqueous Solvents. Comparison of the Thermodynamic Parameters and the Binding of Water to (L-Pro-L-Pro-Gly)_n and (L-Pro-L-Hyp-Gly)_n, *Biopolymers* 16, 601-622.
 95. Sorbetti-Guerri, F., and Michel, D. (1998) Code for Collagen's Stability Deciphered, *Nature* 392, 666-667.
 96. Persikov, A. V., Ramshaw, J. A. M., and Brodsky, B. (2005) Prediction of Collagen Stability from Amino Acid Sequence, *J. Biol. Chem.* 280, 19343-19349.
 97. Perret, S., Merle, C., Bernocco, S., Berland, P., Garrone, R., Hulmes, D. J. S., Theisen, M., and Ruggiero, F. (2001) Unhydroxylated Triple Helical Collagen I Produced in Transgenic Plants Provides New Clues on the Role of Hydroxyproline in Collagen Folding and Fibril Formation, *J. Biochem.* 276, 43693-43698.
 98. Berisio, R., Granata, V., Vitagliano, L., and Zagari, A. (2004) Characterization of Collagen-Like Heterotrimers: Implications for Triple-Helix Stability, *Biopolymers* 73, 682-688.
 99. Engel, J. (1987) *Advances in Meat Research*, Vol. 4, van Nostrand Reinhold Company, New York.
 100. DeRider, M. L., Wilkens, S. J., Waddell, M. J., Bretscher, L. E., Weinhold, F., Raines, R. T., and Markley, J. L. (2002) Collagen Stability: Insights from NMR Spectroscopic and Hybrid Density Functional Computational Investigations of the Effect of Electronegative Substituents on Prolyl Ring Conformations, *J. Am. Chem. Soc.* 124, 2497-2505.

101. Rele, S., Song, Y., Apkarian, R. P., Qu, Z., Conticello, V. P., and Chaikof, E. L. (2007) D-Periodic Collagen-Mimetic Microfibers, *J. Am. Chem. Soc.* *129*, 14780-14787.
102. Mogilner, I. G., Ruderman, G., and Grigera, J. R. (2002) Collagen Stability, Hydration and Native State, *J. Mol. Graphics Modell.* *21*, 209-213.
103. Klein, T. E., and Huang, C. C. (1998) Computational Investigations of Structural Changes Resulting from Point Mutations in a Collagen-Like Peptide, *Biopolymers* *49*, 167-183.
104. Stultz, C. M. (2002) Localized Unfolding of Collagen Explains Collagenase Cleavage Near Imino-poor Sites, *J. Mol. Biol.* *319*, 997-1003.
105. Radmer, R. J., and Klein, T. E. (2004) Severity of Osteogenesis Imperfecta and Structure of a Collagen-like Peptide Modeling a Lethal Mutation Site, *Biochemistry* *43*, 5314-5323.
106. Stultz, C. M. (2006) The Folding Mechanism of Collagen-Like Model Peptides Explored Through Detailed Molecular Simulations, *Protein Sci.* *15*, 2166-2177.
107. Paci, E., and Karplus, M. (1999) Forced Unfolding of Fibronectin Type 3 Modules: An Analysis by Biased Molecular Dynamics Simulations, *J. Mol. Biol.* *288*, 441-459.
108. Saccá, B., Renner, C., and Moroder, L. (2002) The Chain Register in Heterotrimeric Collagen Peptides Affects Triple Helix Stability and Folding Kinetics, *J. Mol. Biol.* *324*, 309-318.
109. Nerenberg, P. S., and Stultz, C. M. (2008) Differential Unfolding of $\alpha 1$ and $\alpha 2$ Chains in Type I Collagen and Collagenolysis, *J. Mol. Biol.* *382*, 246-256.
110. Orgen, J. P. R. O., Irving, T. C., Miller, A., and Wess, T. J. (2006) Microfibrillar Structure of Type I Collagen in Situ, *Proc. Natl. Acad. Sci. USA* *103*, 9001-9005.
111. Tsai, M. I.-H., Xu, Y., and Dannenberg, J. J. (2005) Completely Geometrically Optimized DFT/ONIOM Triple-Helical Collagen-like Structures Containing the ProProGly, ProProAla, ProProDAla, and ProProDSer Triads, *J. Am. Chem. Soc.* *127*, 14130-14131.
112. Lam, J. S. W., Koo, J. C. P., Hudáky, I., Varro, A., Papp, J. G., Penke, B., and Csizmadia, I. G. (2003) Predicting the Conformational Preferences of N-acetyl-4-hydroxy-L-proline-N0-methylamide from the Proline Residue, *J. Mol. Struct. THEOCHEM* *666-667*, 285-589.
113. Benzi, C., Improta, R., Scalmani, G., and Barone, V. (2002) Quantum Mechanical Study of the Conformational Behavior of Proline and 4R-Hydroxyproline Dipeptide Analogues in Vacuum and in Aqueous Solution, *J. Comput. Chem.* *23*.
114. Mooney, S. D., Kollman, P. A., and Klein, T. E. (2002) Conformational Preferences of Substituted Prolines in the Collagen Triple Helix, *Biopolymers* *64*, 63-71.
115. Kawahara, K., Nishi, Y., Nakamura, S., Uchiyama, S., Nishiuchi, Y., Nakazawa, T., Ohkubo, T., and Kobayashi, Y. (2005) Effect of Hydration on the Stability of the Collagen-like Triple-Helical Structure of [4(R)-Hydroxyprolyl-4(R)-hydroxyprolyl]glycine]₁₀, *Biochemistry* *44*, 15812-15822.
116. Mitomo, D., Watanabe, Y. S., Kamiya, N., and Higo, J. (2006) Explicit and GB/SA solvents: Each with Two Different Force Fields in Multicanonical Conformational Sampling of a 25-residue Polypeptide, *Chem. Phys. Lett.* *427*, 399-403.

117. Grigoriev, F. V., Basilevsky, M. V., Gabin, S. N., Romanov, A. N., and Sulimov, V. B. (2007) Cavitation Free Energy for Organic Molecules Having Various Sizes and Shapes, *J. Phys. Chem. B* *111*, 13748-13755.
118. Höfinger, S., and Zerbetto, F. (2003) On the Cavitation Energy of Water, *Chem. Eur. J.* *9*, 566-569.
119. Brown, R. A., and Case, D. A. (2006) Second Derivatives in Generalized Born Theory, *J. Comput. Chem.* *27*, 1662-1675.
120. (2010) NIST Computational Chemistry Comparison and Benchmark Database, (III, R. D. J., Ed.).
121. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations, *J. Comput. Chem.* *14*, 1999.
122. Jorgensen, W. L., Chandrasekhar, J., 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.
123. Darden, T., York, D., and Pedersen, L. (1993) Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems, *J. Chem. Phys.* *98*, 10089-10092.
124. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes, *J. Comput. Phys.* *23*, 327-341.
125. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998) Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices *J. Am. Chem. Soc.* *120*, 9401 -9409.
126. (2007) Multiscale Modeling Tools for Structural Biology
127. Merz, K. M. (2010) Limits of Free Energy Computation for Protein-Ligand Interactions, *J. Chem. Theory Comput.* *6*, 1769-1776.
128. Schäfer, A., Horn, H., and Ahlrichs, R. (1992) Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr, *J. Chem. Phys.* *97*, 2571-2577.
129. Schäfer, A., Klamt, A., Sattel, D., Lohrenz, J. C. W., and Eckert, F. (2000) COSMO Implementation in TURBOMOLE: Extension of an Efficient Quantum Chemical Code Towards Liquid Systems, *Phys. Chem. Chem. Phys.* *2*, 2187-2193.
130. Perdew, J. P., Burke, K., and Ernzerhof, M. (1996) Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.* *77*.
131. Elstner, M., Hobza, P., Suhai, S., and Kaxiras, E. (2001) Hydrogen Bonding and Stacking Interactions of Nucleic Acid Base Pairs: A Density-Functional-Theory Based Treatment, *J. Chem. Phys.* *114*, 5149-5155.