UNIVERSIDAD NACIONAL DEL SUR

TESIS

DOCTOR EN MATEMATICA

El problema del plegamiento de una proteína: de un modelo estocástico a una solución semiempírica

Andrés Colubri

BAHIA BLANCA **ARGENTINA**

2001

DIRECTOR DE TESIS

Ariel Fernández, Ph.D.

Departamento de Matemática e Instituto de Matemática de Bahía Blanca

(INMABB, UNS-CONICET), Universidad Nacional del Sur

Prefacio

Esta tesis es presentada como parte de los requisitos para optar al grado de Doctor en Matemática de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el Instituto y Departamento de Matemática, ante el cual se presenta esta tesis, durante el período comprendido entre el año 1998 y Septiembre de 2001, bajo la dirección del Dr. Ariel Fernández, Profesor Titular del Departamento de Matemática.

Esta tesis ha sido desarrollada en el Instituto de Matemática, dependiente de la Universidad Nacional del Sur y del Consejo Nacional de Investigaciones Científicas y Técnicas, y en el Departamento de Matemática de la Universidad Nacional del Sur. A estas instituciones les agradezco el apoyo material brindado para la concreción de este trabajo. Agradezco también a la Universidad Nacional del Sur y al Consejo Nacional de Investigaciones Científicas y Técnicas por las becas otorgadas que me permitieron desarrollar esta tesis. Agradezco especialmente a mi director, Dr. Ariel Fernández, por enseñarme, guiarme y aconsejarme durante todos estos años de trabajo en común.

Andrés Colubri

Noviembre, 2001.

DEPARTAMENTO DE MATEMÁTICA

UNIVERSIDAD NACIONAL DEL SUR

Dedicada a familia, en especial a mi madre.

Publicaciones suscitadas por la tesis

Las investigaciones que condujeron a la presente tesis también dieron origen a los siguientes artículos científicos, todos ellos publicados o enviados a publicación en revistas especializadas de nivel internacional:

[1] Topology to geometry in protein folding: β-lactoglobulin. Ariel Fernández, Andrés Colubri, R. Stephen Berry. *Proceedings of the National Academy of Sciences, USA* **97**, 14062-14066 (2000).

[2] Topologies to geometries in protein folding: hierarchical and non-hierarchical scenarios. Ariel Fernández, Andrés Colubri, R. Stephen Berry. *Journal of Chemical Physics* **114**, 5871-5887 (2001).

[3] Renormalized Hamiltonian for a peptide chain: Digitalizing the protein folding problem. Ariel Fernández, Andrés Colubri. *Journal of Mathematical Physics* **41**, 2593-2603 (2000).

[4] Microscopic dynamics from a coarsely defined solution to the protein folding problem. Ariel Fernández, Andrés Colubri. *Journal of Mathematical Physics* **39**, 3167-3187 (1998).

[5] Semiempirical prediction of protein folds. Ariel Fernández, Gustavo Appignanesi, Andrés Colubri. *Physical Review E* **64**. 21901-21914 (2001).

[6] Computational dissection of ultra-fast events in protein folding. Tobin R. Sosnick, R. Stephen Berry, Andrés Colubri, Ariel Fernández. *Proceedings of the National Academy of Sciences, USA*. Enviado para su publicación.

[7] Three bodies correlations in protein folding: the origin of cooperativity. Ariel Fernández, Andrés Colubri, R. Stephen Berry. Aceptado para su publicación en *Physica A*.

[8] Finding the collapse-inducing nucleus in a folding protein. Ariel Fernández, Andrés Colubri, Gustavo Appignanesi. *Journal of Chemical Physics* **114**, 8678-8684 (2001).

[9] Coarse semiempirical solution to the protein folding problem. Ariel Fernández, Andrés Colubri, Teresita Burastero, Gustavo Appignanesi. *Physica A* **293**, 358-384 (2001).

[10] Nucleation theory for helix unfolding in peptide chains. Ariel Fernández, Andrés Colubri. *Physical Review E* **60**, 4645-4651 (1999).

[11] How large should proteins be?: The minimal size of a good structure seeker. Ariel Fernández, Andrés Colubri, Teresita Burastero, Ana Tablar. *Phys. Chem. Chem. Phys*. **1**, 4347-4354 (1999).

[12] Semiempirical variational approach to RNA folding. Ariel Fernández, Andrés Colubri. *Physica A* **248**, 336-352 (1998).

Índice general

Resumen y organización de la tesis

El problema del plegamiento de proteínas consiste en predecir la conformación nativa (funcional) de una proteína con el sólo conocimiento de su secuencia primaria (composición química). Este constituye un problema central y abierto en la biología molecular, no solo por su significado teórico en el campo de la matemática aplicada sino también por sus consecuencias prácticas en el terreno de la biotecnología.

Los métodos "clásicos" de la física-matemática que requieren la incorporación de todos los detalles conformacionales del sistema no han conducido a la solución de este problema, fundamentalmente porque una proteína es un sistema muy complejo cuando es descripta en todo su detalle. Las ecuaciones del movimiento que resultan de aplicar estos métodos son imposibles de resolver analíticamente, y más aún, su solución numérica está muy lejos del alcance de las computadoras mas potentes de la actualidad.

Sin embargo, se ha corroborado experimentalmente que el proceso de plegamiento posee tres características que hemos decidido destacar: **(a)** es expeditivo, es decir, ocurre en escalas de tiempo inconmensurablemente más cortas que las escalas ergódicas o termodinámicas; **(b)** es robusto, ya que tolera cambios importantes en las condiciones del ambiente circundante o en la misma proteína; y **(c)** es altamente reproducible. Estas observaciones sugieren que la dinámica del plegamiento no depende de los detalles conformacionales más finos de una proteína, sino que, por el contrario, es posible encontrar una representación simplificada que baste para reproducir las características esenciales del plegamiento, y en particular, que permita resolver el problema central indicado al principio.

En vista de lo dicho, los objetivos centrales de esta tesis son los siguientes: **(a)** utilizar los elementos empíricos descriptos recién para construir un modelo teórico del plegamiento en el cual las conformaciones de una proteína estén representadas de manera simplificada y discreta, compromiso aparentemente indispensable para abarcar computacionalmente tiempos reales de plegamiento; **(b)** definir una dinámica markoviana en el espacio de conformaciones simplificadas, basada en una función potencial que sea tratable numéricamente; **(c)** probar que esta dinámica corresponde rigurosamente a la proyección del proceso de plegamiento que se lleva a cabo en el espacio de conformaciones original; y **(d)** elaborar un algoritmo que permita efectivamente generar caminos de plegamiento de proteínas naturales con un grado de resolución estructural mínimo pero suficientemente revelador.

El algoritmo de plegamiento ha sido exitosamente implementado en un programa bautizado "Máquina de Plegar" o "Folding Machine" (FM), que se ejecuta en una computadora personal. Con él hemos generado trayectorias de plegamiento que están en excelente acuerdo con observaciones experimentales. En esta tesis se estudian en detalle las trayectorias de plegamiento de dos proteínas particulares: ubiquitina y la variante hipertermófila de la proteína G. Ambas tienen una estructura nativa similar y comparten la misma topología, pero sin embargo difieren enormemente en el grado de diversidad de caminos de plegamiento.

La tesis está organizada de la siguiente manera:

Capítulo 1: Este capítulo es esencialmente de carácter físico-químico. A efectos de ilustrar la importancia que tiene la solución del problema del plegamiento, se indica el rol esencial que juegan las proteínas en la naturaleza. Luego se describe en detalle las características conformacionales de una proteína y se muestra como se pueden introducir reducciones importantes en el espacio conformacional, sin llegar todavía al concepto de discretización, el cual será tratado en el siguiente capítulo. Se plantea con detenimiento el problema central del plegamiento, junto con otros problemas derivados que también son tratados en la tesis. Se finaliza mencionado algunos antecedentes importantes en el estudio del problema a efectos de ejemplificar el fracaso de las técnicas convencionales para generar trayectorias de plegamiento.

Capítulo 2: Este capítulo constituye la parte matemática de la tesis. Comienza por describir los modelos matemáticos que se utilizan usualmente para describir el plegamiento. A continuación se expone la motivación y la justificación de la simplificación conceptual que supone una discretización en el espacio de conformaciones de una proteína. Una vez definido el nuevo espacio discreto, se muestra como se construye en él un proceso markoviano que representa rigurosamente a una proyección realista de la dinámica del plegamiento. Aquí se introducen conceptos y resultados matemáticos que son analizados en detalle, puesto que son novedosos en el campo de los procesos estocásticos y resultan ser fundamentales en nuestro modelo. Al final del capítulo, estamos en condiciones de presentar un esquema general del algoritmo de plegamiento, ejecutable en una computadora personal.

Capítulo 3: En este capítulo se expone detenidamente otro elemento, que si bien no es imprescindible para brindar una exposición general de nuestro modelo, es esencial a efectos de llegar a una implementación computacional concreta: la función potencial que caracteriza las distintas interacciones en una proteína. Las probabilidades de transición del proceso markoviano discreto introducido en el capítulo anterior están definidas precisamente en términos de este potencial. Las razones físicas que justifican cada uno de los términos que constituyen este potencial son expuestas.

Capítulo 4: El algoritmo de plegamiento, anticipado al final del capítulo 2, es ahora descripto en todo detalle.

Capítulo 5: En este capítulo se ilustra el poder predictivo del algoritmo, aplicándolo al plegamiento *in-vitro* de proteínas específicas.

Capítulo 1

Las proteínas y su proceso de plegamiento

1.1 Las proteínas y su proceso de plegamiento

Con el único propósito de introducir al lector no familiarizado en el problema del plegamiento de proteínas y brindarle una imagen clara del sistema que estudiamos en esta tesis, la siguiente es una aproximación válida: podemos pensar en una molécula de proteína como un conjunto de N "cuentas" o esferas rígidas unidas entre sí de manera lineal por enlaces de longitud fija (Fig. 1). La proteína es libre de girar alrededor de cada uno de estos enlaces, los cuales definen en consecuencia N-1 grados de libertad angulares o torsionales. Aparte de las N cuentas que forman la molécula de proteína propiamente dicha, hay que considerar una enorme cantidad M de "cuentas" adicionales que rodean a la proteína y que se mueven libremente en el espacio. Estas cuentas constituyen el solvente en el cual la proteína está inmersa. Sobre cada una de las cuentas actúan distintas fuerzas que se originan en interacciones mutuas de atracción o repulsión, que se pueden deber, por ejemplo, a que algunas cuentas tengan carga eléctrica. Por otra parte, el sistema proteína-solvente es conservativo, por lo que existe una función potencial U de la cual podemos derivar todas las fuerzas de interacción.

Debido a la acción de estas fuerzas, la molécula de proteína y el solvente circundante efectuarán movimientos en el espacio que los llevarán, desde una conformación inicial dada, a una estructura final de cierta estabilidad que corresponda a un mínimo del potencial (que no necesariamente es el mínimo global). En este momento, podemos esbozar al problema del plegamiento de la siguiente manera: dado un cierto estado inicial y conocido el tipo de "cuentas"

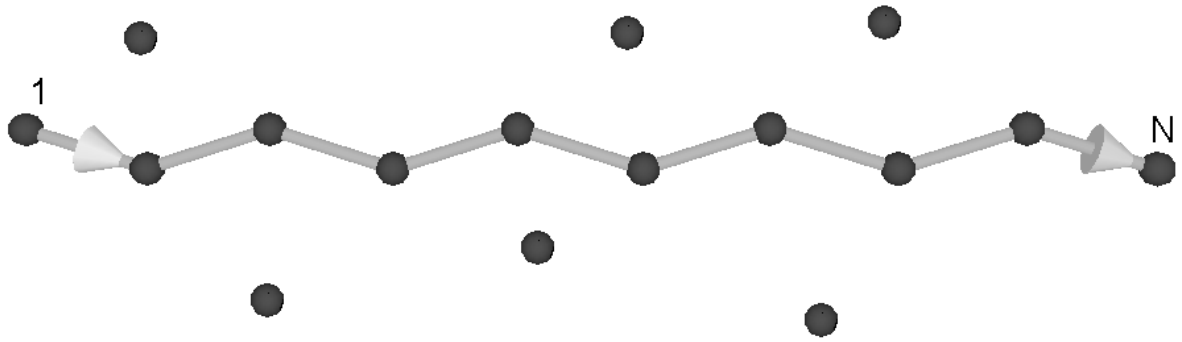


Figura 1: Esquema de una proteína y el solvente circundante. La proteína está representada por un conjunto de "cuentas" o esferas rígidas, conectadas linealmente entre sí. Las esferas libres alrededor de la proteína constituyen el solvente en el cual está inmersa.

que forman al sistema (cuáles son sus cargas eléctricas, masas, etcétera) ¿cuál es la trayectoria que éste lleva a cabo? En particular, ¿cuales son las conformaciones espaciales adoptadas en cierto intervalo de tiempo? El término plegamiento se debe simplemente a que las estructuras que adopta una proteína real son generalmente muy compactas y presentan numerosos "dobleces" o "plegamientos" sobre sí misma. La palabra plegamiento es utilizada inclusive como sinónimo de estructura o conformación espacial.

1.2 Importancia del problema

La descripción extremadamente simplificada de una proteína y del problema de plegamiento expuesta en la sección anterior sugiere que, en principio, estamos ante un problema de n cuerpos típico: se conoce la función potencial del sistema y nuestro objetivo consiste en determinar su evolución temporal. Las preguntas que podemos hacernos ahora son las siguientes: ¿qué justifica resolver este problema de n cuerpos particular? ¿es relevante desde el punto de vista de la matemática aplicada?

En primer lugar, las proteínas tienen una importancia intrínseca debido a que son moléculas con una relevancia biológica fundamental. Las proteínas intervienen en la mayor parte de las funciones que mantienen la vida [1]:

(a) Catálisis de las reacciones bioquímicas. Casi todas las reacciones químicas de las células están catalizadas por proteínas específicas (llamadas enzimas), y serían imposibles de realizar en ausencia de las mismas.

(b) Defensa celular: Las proteínas llamadas anticuerpos o inmunoglobulinas son fundamentales en el mecanismo de defensa contra agentes infecciosos externos.

(c) Estructurales: Algunas proteínas estructurales están presentes en el interior de las células, formando filamentos o láminas que determinan la su forma y tamaño. Otras son extracelulares y unen las células de varias maneras para fortalecer los tejidos u órganos y los proveen de una matriz en la cual pueden crecer.

(d) Regulación del código genético: Existen proteínas que intervienen en la replicación del ADN, otras participan en la reparación de partes dañadas del material genético. Muchos de estos procesos tienen vinculación directa con el origen de enfermedades genéticas, el cáncer y el envejecimiento celular.

Lo expuesto recién brinda una somera idea acerca de la importancia fundamental de las proteínas en el ámbito biológico. Resulta claro de esta breve introducción que un modelo teórico que permita describir adecuadamente la función de las proteínas tendría enormes consecuencias en el programa investigativo de la era post-genómica, entre las cuales figuran importantes avances en el campo de la biotecnología y la medicina [2-4].

Por otra parte, es bien conocido que la función de una proteína natural está unívocamente determinada por la estructura espacial adoptada durante el proceso de plegamiento [1, 5], la cual

es denominada por este motivo estructura funcional o nativa. Este hecho confiere de una relevancia aún mayor al estudio teórico del plegamiento, ya que el mismo nos permitiría acceder a una mayor comprensión de los mecanismos por los cuales una proteína es capaz de llevar a cabo sus funciones biológicas.

Desde el punto de vista de la matemática, el problema del plegamiento propone nuevos desafíos cuya superación puede conducir al desarrollo de nuevas ideas y herramientas teóricas. Como veremos más adelante, la aplicación de los métodos tradicionales de la física-matemática, esto es, plantear la 2da Ley de Newton y luego resolver la ecuación diferencial asociada es totalmente inútil en el contexto del plegamiento, ya sea que intentemos resolver el problema analíticamente o numéricamente. Esta dificultad se debe básicamente a que el número de cuerpos en el sistema es tremendamente alto. En consecuencia, tendremos que ser capaces de introducir simplificaciones drásticas en el plano representacional, pero sin perder las características físicas esenciales que determinan el plegamiento. Tal es la motivación central de esta tesis. Para poder hacer esto empezaremos, en la siguiente sección, por describir detalladamente a las proteínas desde el punto de vista estructural.

1.3 La estructura espacial de una proteína

Una proteína está formada por unidades constitutivas básicas denominadas aminoácidos, las cuales se conectan linealmente entre sí formando una cadena altamente flexible. Existen numerosos aminoácidos que se diferencian entre sí por sus características físico-químicas: tamaño, afinidad con el solvente, carga eléctrica, etcétera [1, 5]. Veinte aminoácidos se destacan sobre los demás por ser los más frecuentes en las proteínas naturales. Estos 20 aminoácidos "normales" son mostrados en la Tabla I. Muchos de los aminoácidos "atípicos" restantes se

obtienen a partir de alteraciones específicas en los aminoácidos normales [5]. A lo largo de esta tesis sólo utilizaremos a los 20 aminoácidos típicos.

La afinidad con el solvente es un parámetro particularmente relevante en la caracterización de los aminoácidos, y de acuerdo al mismo los 20 aminoácidos pueden clasificarse en 4 grupos (ver Tabla I) [5]:

(a) Hidrofóbicos: Su interacción con el solvente es de carácter repulsivo. Es por esto que en la estructura nativa tienden a ocupar las regiones internas de la proteína, protegidas del solvente (el llamado núcleo hidrofóbico).

(b) Polares: Interactúan con el solvente de manera atractiva. Al contrario de los aminoácidos hidrofóbicos, muestran preferencia a ubicarse en las regiones externas de la proteína, expuestas al solvente.

(c) Amfifílicos: Son aminoácidos que tienen una parte hidrofóbica y otra polar, por lo cual pueden ocupar indistintamente las zonas internas y externas de la proteína.

(c) Indiferentes: Carecen de una preferencia particular por el solvente.

Diremos que una proteína tiene longitud N si está formada por N aminoácidos. Se define la secuencia primaria de una proteína a la sucesión de aminoácidos $s = (a_1, a_2, \ldots, a_N)$, donde N es la longitud de la proteína y a_i es su i-ésimo aminoácido. En las referencias es frecuente que los términos unidad o residuo se utilizen en lugar de aminoácido, convención que seguiremos en esta tesis. La composición química de una proteína queda totalmente especificada por su secuencia primaria.

Los 20 aminoácidos comparten un patrón estructural similar: todos ellos están formados por un grupo $^{+}H_3N$ (grupo amida), unido a un grupo CHR, el cual está unido a su vez a un grupo COO^{-} (grupo carbonilo). Aquí H, N, C y O denotan a los átomos de hidrógeno, nitrógeno,

carbono y oxígeno, respectivamente, mientras que R es un grupo variable de átomos llamado residuo o cadena lateral. La fórmula química de un aminoácido genérico es (ver Fig. 2-a):

$$A = {}^{+}H_3N—CHR—COO^{-}$$

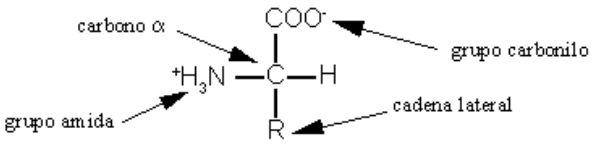
Alanina	A	ALA	hidrofóbico
Arginina	R	ARG	polar
Asparagina	N	ASN	polar
Aspartato	D	ASP	polar
Cisteína	C	CYS	amfifílico
Fenilalanina	F	PHE	hidrofóbico
Glutamina	Q	GLN	polar
Glutamato	E	GLU	polar
Glicina	G	GLY	indiferente
Histidina	H	HIS	polar
Isoleucina	T	ILE	hidrofóbico
Leucina	L	LEU	hidrofóbico
Lisina	K	LYS	polar
Metionina	M	MET	hidrofóbico
Prolina	P	PRO	hidrofóbico
Serina	S	SER	polar
Tirosina	Y	TYR	amfifílico
Treonina	T	THR	polar
Triptofano	W	TRP	hidrofóbico
Valina	V	VAL	hidrofóbico

Tabla I: Los 20 aminoácidos. En la primera columna está indicado el nombre completo, en la segunda el código de una letra y en la tercera, el de tres letras. La cuarta columna indica la afinidad por el solvente de cada aminoácido.

Figura 2 (siguiente página): Estructura genérica de un aminoácido (a) y estructura de cada uno de los 20 aminoácidos (b). Todos son mostrados proyectados en un plano que pasa por el carbono α. En la realidad, el grupo amida (${}^{+}H_3N$) y el hidrógeno están arriba del plano de la hoja, mientras que el grupo carbonilo (COO^{-}) y la cadena lateral están abajo (asumiendo que los aminoácidos son de tipo L). Los aminoácidos están agrupados de acuerdo a su preferencia por el solvente (hidrofóbicos, polares, amfifílicos e indiferentes).

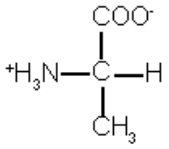
a

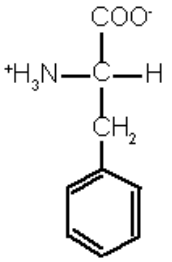
Aminoácido genérico

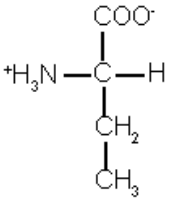


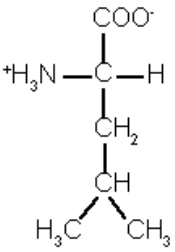
b

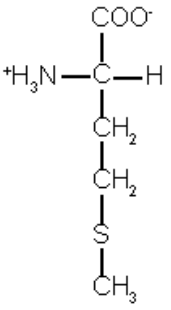
Hidrofóbicos

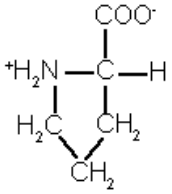

Alanina

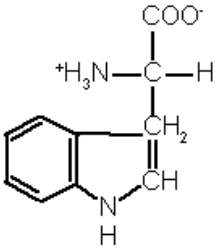

Fenilalanina

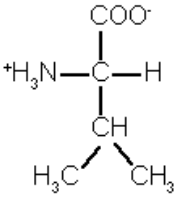

Isoleucina


Leucina

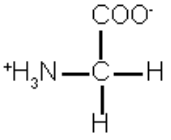

Metionina

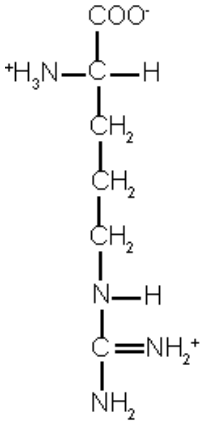

Prolina


Triptofano

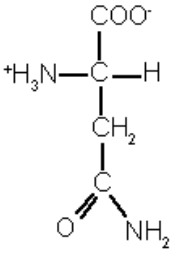

Valina

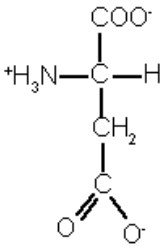
Indiferente

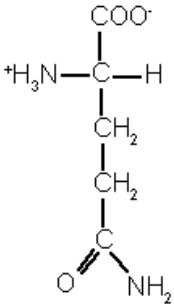

Glicina

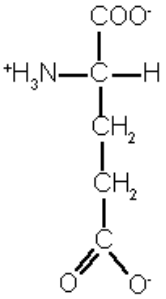

Arginina

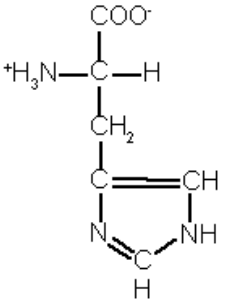
Polares

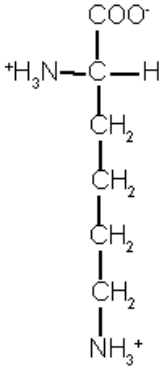

Asparagina

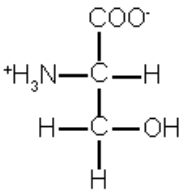

Aspartato

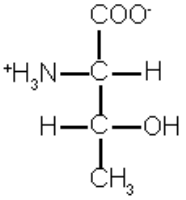

Glutamina


Glutamato

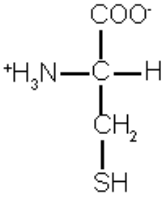

Histidina

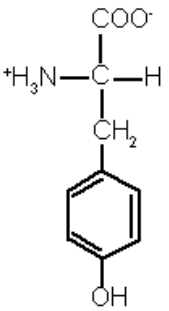

Lisina


Serina


Treonina

Amfifílicos


Cisteína


Tirosina

La identidad particular de cada aminoácido está determinada por la cadena lateral R, en la cual se originan las distintas propiedades físico-químicas que diferencian a los 20 aminoácidos. En la Figura 2-b se muestran las estructuras de los 20 aminoácidos. El carbono del grupo CHR es llamado carbono α (C^{α}). Una inspección de la Figura 2-b muestra que todas las cadenas laterales comienzan en un carbono, excepto en el aminoácido glicina (GLY), en el cual la cadena lateral se reduce a un hidrógeno. Este carbono es designado carbono β (C^{β}). En todos los 20 aminoácidos, excepto la glicina, el C^{α} es un centro asimétrico: el C^{β} puede apuntar hacia un lado o el otro, dando lugar a los aminoácidos L (mostrados en la Figura 2) y los aminoácidos D. En las proteínas sintetizadas naturalmente, los aminoácidos son siempre de de tipo L [5]. Es por esto que esta tesis sólo consideraremos aminoácidos de tipo L.

En una proteína, el aminoácido i se une con el i+1 a través de la formación de un enlace covalente, llamado enlace peptídico, entre el oxígeno i y el nitrógeno i+1. La siguiente fórmula química representa un enlace peptídico entre los aminoácidos i e i+1:

$$—HN_i— C^{\alpha}_i HR_i— C_iO—HN_i— C^{\alpha}_{i+1} HR_{i+1}— C_{i+1}O—$$

En la Figura 3 se muestra la estructura espacial de un fragmento de cadena proteínica. Se puede observar que los átomos N, C^{α} y C de cada aminoácido forman una columna vertebral (backbone) del cual emergen hacia los lados las cadenas laterales.

Los enlaces covalentes simples del backbone, $N_i—C^{\alpha}_i$, $C^{\alpha}_i—C_i$ y $C_i—N_{i+1}$, pueden rotar fácilmente alrededor de su eje. Los ángulos torsionales asociados a cada uno de estos ejes son denominados, respectivamente, ϕ_i, ψ_i y ω_i [5]. Los demás enlaces covalentes simples de la proteína, por ejemplo, los presentes en las cadenas laterales, determinan ángulos torsionales de manera similar. El ángulo ω_i resulta estar fijo en 180° debido a que en realidad el enlace $C_i—N_{i+1}$

es parcialmente un enlace doble, lo cual impide rotaciones alrededor del mismo. Otros ángulos torsionales resultan también ser innecesarios, como por ejemplo el asociado al enlace N_i—H, ya que la proteína es simétrica con respecto a rotaciones alrededor de los mismos.

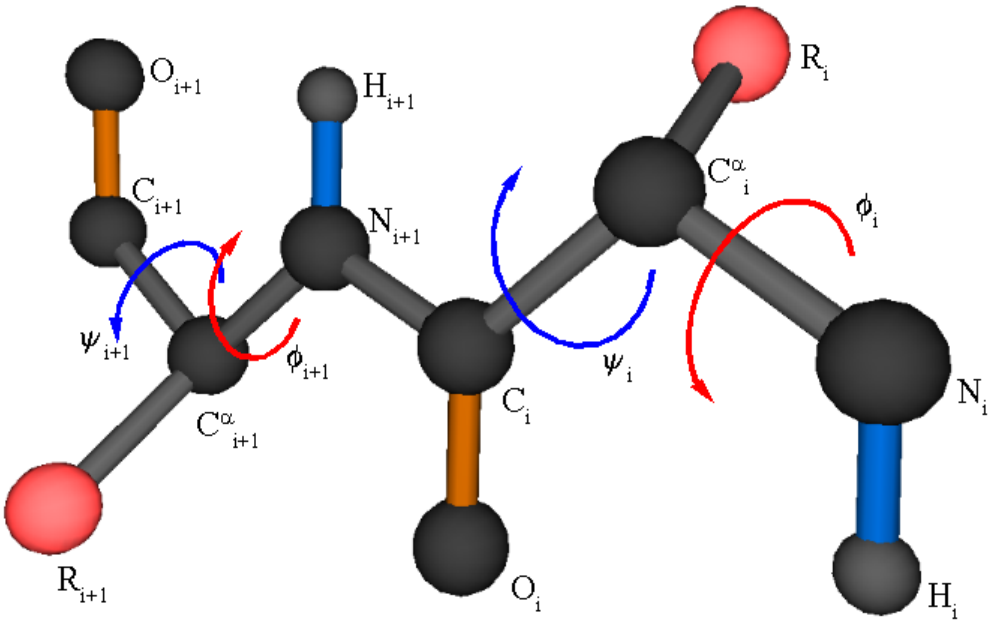


Figura 3: Estructura espacial de un fragmento del backbone formado por dos aminoácidos L consecutivos, correspondiente a la conformación más extendida posible ($\phi = \psi = \omega = 180°$). La cadena lateral está representada genéricamente como una esfera, pero en la realidad corresponde a un grupo más o menos numeroso de átomos. El hidrógeno enlazado al carbono α no es mostrado en la figura.

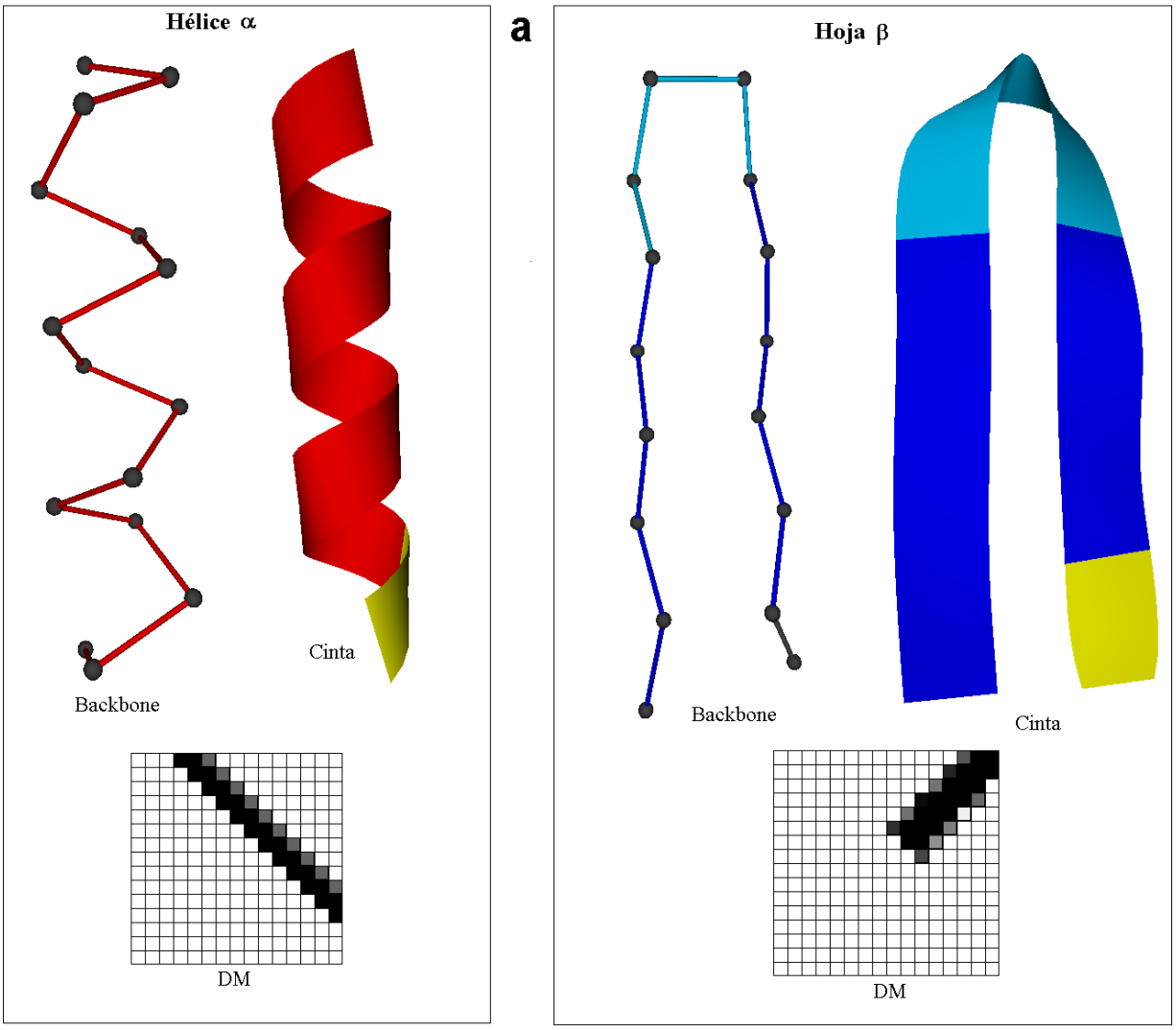
Los ángulos torsionales ϕ_i y ψ_i del backbone son particularmente importantes. Se observa que ciertas combinaciones de ángulos (ϕ_i, ψ_i) en regiones adyacentes de la cadena aparecen en todas las proteínas. Estas conformaciones torsionales típicas determinan estructuras locales bien definidas, llamadas estructuras secundarias [5]. Las estructuras secundarias más frecuentes son la hélice α, la hoja β y el codo β (β-turn), todas ellas mostradas en la Figura 4-a:

(a) La hélice α es una estructura que se obtiene cuando más de 4 aminoácidos consecutivos adoptan valores (ϕ, ψ) próximos a (57°, 47°), o bien a (-57°, -47°). Cada vuelta de la hélice involucra a 3,6 aminoácidos, mientras que el desplazamiento a lo largo del eje de la hélice es de 1,5 Å por aminoácido. Hay dos tipos de hélices α, que se diferencian por su sentido de giro: las dextrógiras (las más comunes y las que presentan los ángulos (57°, 47°)), que giran en sentido horario al desplazarse del extremo N al extremo C del backbone, y las levógiras, que giran en sentido contrario y son muy infrecuentes.

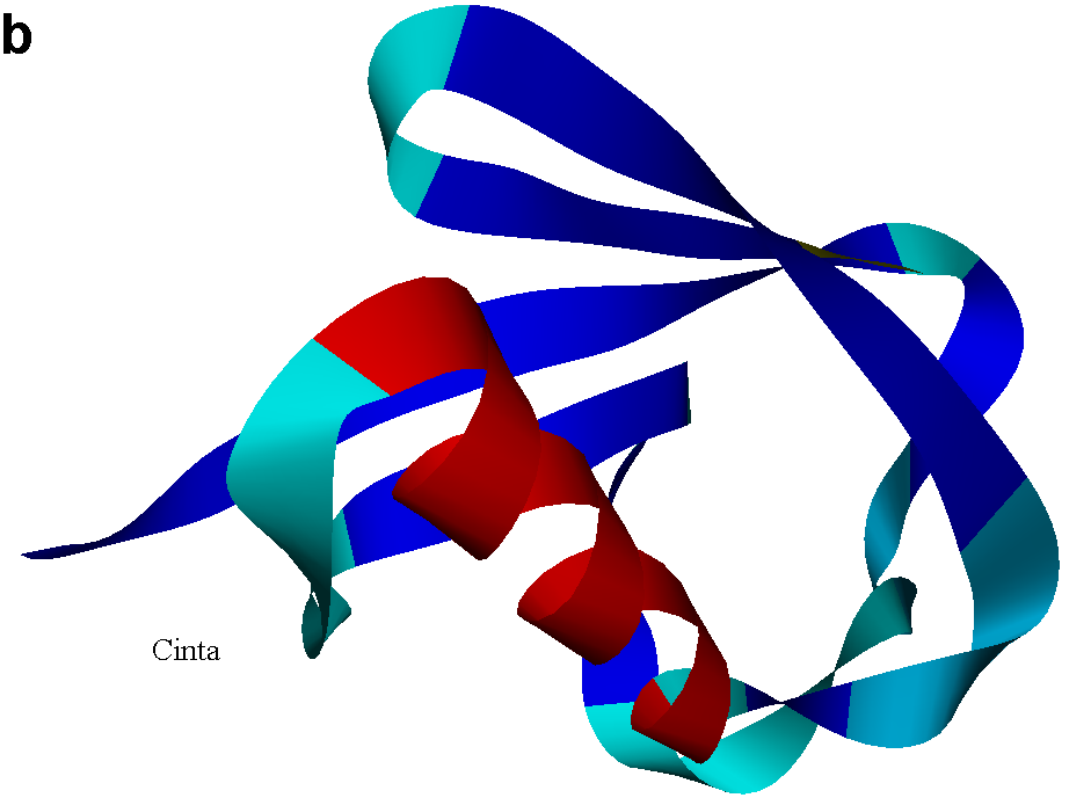
(b) Una hoja β está formada por 2 porciones extendidas de la cadena que se enfrentan de manera paralela o anti-paralela, llamadas hebras β (β-strands). Los valores (ϕ, ψ) en cada unidad de un β-strand son próximos a (-119°, 113°) o bien a (-139°, 135°). La geometría del backbone en los β-strands se aproxima a las conformaciones más extendidas que puede adoptar una cadena proteínica, siendo el desplazamiento a lo largo del eje del strand de 3,47 Å por aminoácido.

(c) Un β-turn consiste en 4 aminoácidos que adoptan la forma de una "U", lo cual permite que la cadena se doble sobre sí misma. Los ángulos (ϕ, ψ) tienen un rango de valores más amplio que en las hélices α y hojas β.

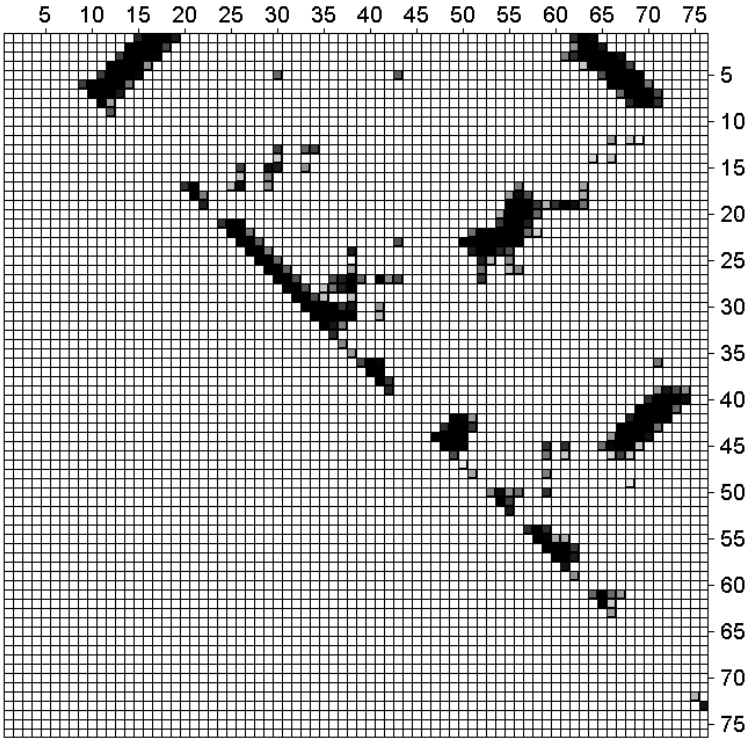
Figura 4 (las dos páginas siguientes): En (a) se muestran las estructuras secundarias típicas: hélice α, β-strand, β-turn y hoja β. En cada caso se utiliza una representación simplificada del backbone en la que sólo se indican los carbonos α, una representación esquemática con forma de cinta, y finalmente la matriz de distancias o DM. En principio, la posición (i, j) de la DM contiene la distancia entre los carbonos α i y j de la cadena. En la figura, las DM's están coloreadas con una escala de grises modulada con esta distancia: la entrada (i, j) es negra para distancias es menores que 7,5 Å, mientras que entradas grises indican distancias entre 7,5 Å y 8,5 Å. Se aprecia que cada estructura secundaria tiene un patrón característico bien definido en la DM. En (b) se muestra la estructura espacial nativa de una proteína natural (1ubi), utilizando la representación de cintas y la DM. Se puede apreciar las distintas estructuras secundarias locales y el patrón de interacciones terciarias de largo rango entre ellas.


Hélice α
Backbone
Cinta
DM
a
Hoja β
Backbone
Cinta
DM
β-turn
Backbone
Cinta
DM

b



Cinta


5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75

DM

A su vez, hélices α y hojas β distantes en la cadena pueden interactuar entre sí cuando son aproximadas por movimientos en las unidades intermedias. Estas regiones que permiten la conexión de estructuras secundarias lejanas generalmente carecen de una geometría (ϕ, ψ) definida y son llamadas lazos (loops). El patrón de interacciones entre las estructuras secundarias locales de una proteína se denomina estructura terciaria.

Es claro que todas las características estructurales recién indicadas quedan totalmente especificadas por la posición de cada uno de los átomos de la proteína. Si N_i es el número de átomos del aminoácido i-ésimo, entonces el número de coordenadas necesario para describir la geometría interna de la proteína es:

$$N_p = 3 \left(\sum_{i=1...N} N_i \right)$$

Sin embargo, para caracterizar las interacciones y la dinámica de una proteína, es necesario incorporar también al solvente. En general, el solvente está formado por moléculas de agua, cada una de las cuales está compuesta por tres átomos. Si M es el número de moléculas de solvente, entonces el número de coordenadas necesario para describir al solvente es:

$$N_s = 9\ M$$

Por lo tanto, el espacio de conformaciones que incorpora todos los detalles del sistema proteína-solvente es $X^0 = R^{N_p+N_s}$. Por ejemplo, para una proteína de tamaño intermedio se tiene $N \approx 150$, y como $Ni \approx 10$, resulta que el número de coordenadas necesarias para caracterizar a la proteína llega a $N_p \approx 4500$. Esto sin embargo, no es todo: para representar adecuadamente al solvente en todo detalle es necesario un enorme número de moléculas, fácilmente se puede tener $M \approx 1000$. En consecuencia, el número total de coordenadas asciende a $N_p + N_s \approx 15000$.

Por otra parte, si ignoramos la geometría de las cadenas laterales y la estructura del solvente, se observa que la conformación espacial de una proteína natural puede ser descripta con bastante precisión en término de las estructuras secundarias que adoptan regiones locales de la cadena y el patrón de interacciones terciarias entre las estructuras secundarias (ver Figura 4-b). A su vez, las estructuras secundarias y terciarias quedan caracterizadas totalmente por los ángulos torsionales del backbone (ϕ_i, ψ_i), con $1 \leq i \leq N$. Es decir que la geometría torsional del backbone puede ser caracterizada por un punto en el espacio $X = T^{2N}$, el toro de dimensión 2N. La pregunta que surge naturalmente es la siguiente: ¿es posible caracterizar la dinámica torsional sin necesidad de salir de T^{2N}? En principio, no podemos responder afirmativamente a esta pregunta, ya que puede suceder que el movimiento de los ángulos torsionales (ϕ, ψ) dependa de detalles conformacionales más finos, tales como la estructura detallada del solvente o la posición de los átomos de las cadenas laterales. A continuación veremos las razones físicas que justifican una respuesta afirmativa.

1.4 Simplificaciones en el espacio conformacional de una proteína

Como se explicó en la sección anterior, el espacio de conformaciones detalladas del sistema proteína-solvente es $X^0 = R^{N_p+N_s}$. La dinámica de este sistema queda caracterizada por una función potencial U: $X^0 \rightarrow R$. Nuestro objetivo es mostrar que las conformaciones del sistema y la dinámica asociada pueden ser simplificados hasta llegar al espacio $X = T^{2N}$, donde un potencial efectivo U^{eff}: $X \rightarrow R$ bastará para caracterizar la dinámica torsional de una proteína. En el siguiente esquema indicamos los pasos que seguiremos para ir desde $[X^0 = R^{N_p+N_s}, U]$ a $[X = T^{2N}, U^{eff}]$:

$$[X^0 = R^{N_p+N_s},\ U]$$
$$\Big|$$
Eliminación de los grados de libertad de alta frecuencia
$$\downarrow$$
$$[X^1, U^1]$$
$$\Big|$$
Tratamiento implícito del solvente
$$\downarrow$$
$$[X^2, U^2]$$
$$\Big|$$
Simplificación de las cadenas laterales
$$\downarrow$$
$$[X = T^{2N}, U^{eff}]$$

El par $[X^i, U^i]$ representa el espacio conformacional intermedio X^i obtenido al eliminar ciertas coordenadas en el nivel anterior, junto con la función potencial U^i: $X^i \rightarrow R$ que genera la dinámica en dicho nivel representacional.

1.4.1 Eliminación de los grados de libertad de alta frecuencia

Sean $\mathbf{r}_i$ y $\mathbf{r}_j$ los vectores posición de dos átomos de la proteína unidos covalentemente (por ejemplo, N_i y C^{α}_i). La longitud del enlace se define como:

$$r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$$

Si $\mathbf{r}_k$ corresponde a un tercer átomo enlazado covalentemente con j, el ángulo plano determinado por los tres átomos (este es el caso de los átomos del backbone N_i, C^{α}_i y C_i) se define como el ángulo entre los vectores $\mathbf{r}_i - \mathbf{r}_j$ y $\mathbf{r}_k - \mathbf{r}_j$, es decir:

$$\alpha_{ijk} = \arccos[\ (r_{ij}\, r_{jk})^{-1}\ (\mathbf{r}_i - \mathbf{r}_j)\cdot(\mathbf{r}_k - \mathbf{r}_j)\ ]$$

Se observa experimentalmente que las coordenadas r_{ij} y α_{ijk} oscilan muy rápidamente alrededor de valores de equilibrio fijos [6]. Dado que: **(a)** el período de estas oscilaciones es muy

corto (10^{-15}-10^{-12} s) en comparación con las escalas de tiempo en las cuales se desea observar la dinámica y **(b)** la amplitud de estas oscilaciones es pequeña, ya que grandes distorsiones respecto a los valores de equilibrio son muy penalizadas energéticamente por el potencial U, podemos introducir la siguiente simplificación: las longitudes de enlaces y los ángulos planos son constantes en el tiempo, estando fijos en sus respectivos valores de equilibrio.

Esta simplificación nos permite describir completamente a la geometría interna de la proteína en función de sus ángulos torsionales solamente.

1.4.2 Tratamiento implícito del solvente

La descripción detallada de las moléculas de solvente es el principal motivo por el cual la dimensionalidad del sistema es tan elevada. Por consiguiente, una representación implícita de las interacciones proteína-solvente [7] es imprescindible para obtener un marco representacional más sencillo y manejable. Por supuesto que una simplificación tan drástica debe estar adecuadamente justificada en términos físicos.

En particular, si las siguientes aproximaciones son válidas: **(a)** la energía interna del solvente es constante, y **(b)** la interacción del solvente con la proteína queda determinada exclusivamente por la conformación de la proteína, entonces es posible eliminar el tratamiento explícito del solvente. La aproximación **(b)** es correcta, en particular, si el solvente puede ser modelado como un contínuo que llena uniformemente todas las regiones del espacio que no están ocupadas por la proteína. En este modelo, la conformación del solvente sólo depende de su volumen, y el mismo está unívocamente determinado por la forma espacial de la proteína. Este tratamiento contínuo del solvente ha sido aplicado exitosamente [8] para calcular las energías de

solvatación en estructuras nativas, lo cual respalda su validez. Por otro lado, la aproximación **(a)** es cierta ya que en las condiciones en las que ocurre el plegamiento, el solvente es incapaz de adoptar conformaciones preferenciales. A excepción de la interfase proteína-solvente, las moléculas del solvente se encuentran permanentemente en un estado de agitación térmica a energía constante [6].

La incorporación implícita del solvente tiene dos consecuencias importantes en la construcción de una función potencial que sólo dependa de las conformaciones internas de la proteína:

(a) El potencial deberá contener nuevos términos efectivos que representen el efecto neto de las interacciones entre los aminoácidos hidrofóbicos y polares y el solvente [9-11].

(b) Las interacciones intramoleculares que dependan explícitamente del nivel de exposición al solvente de los aminoácidos deberán ser modificadas. El efecto explícito del solvente será reemplazado por un parámetro que modele el nivel de exposición como función de la conformación local de la cadena.

Los detalles de la construcción de este potencial efectivo que incorpora implícitamente al solvente serán expuestos en el capítulo 3.

1.4.3 Simplificación de las cadenas laterales

La última simplificación está fundamentada en la siguiente observación: la exploración torsional de las cadenas laterales ocurre en una escala temporal mucho más rápida que la dinámica del backbone. Esto justifica el tratamiento de los movimientos torsionales de las cadenas laterales a través de una representación de las mismas como elipsoides rígidos [9]. La región que determina cada uno de estos elipsoides rígidos es el promedio espacial de las

conformaciones que la cadena lateral es capaz de visitar en tiempos muy cortos, en comparación con las escalas asociadas a los movimientos relativamente lentos de los ángulos ϕ y ψ.

El efecto de esta nueva simplificación en la construcción de la función potencial se traduce en la necesidad de incluir términos entrópicos que representen la pérdida conformacional que sufren las cadenas laterales cuando el espacio que pueden explorar es restringido por la formación de vínculos o interacciones de corto rango. Este aspecto será tratado detenidamente en el capítulo 4.

De esta manera, hemos llegado a que el espacio de conformaciones en el cual se puede estudiar la dinámica de una proteína es $X = T^{2N}$, para lo cual debemos construir un potencial efectivo U^{eff} que incorpore los efectos de todas las simplificaciones efectuadas. El problema de construir U^{eff} no es para nada de importancia secundaria [13]. Desde un punto de vista formal, U^{eff} está vinculado con el potencial riguroso U a través de los sucesivos pasos de reducción o proyección conformacional que hemos descripto. Sin embargo, la complejidad intrínseca de la función U vuelve imposible encontrar un método analítico para obtener directamente U^{eff} a partir de U. En lugar de esto, debemos recurrir a una combinación de elementos empíricos y analíticos que nos conduzca a una función U^{eff} que represente correctamente las interacciones determinantes de la dinámica torsional del backbone.

1.5 Restricciones del movimiento: las cuencas de Ramachandran

Se observa que para cada par de coordenadas torsionales (ϕ_i, ψ_i) no todos los valores en [-180°, 180°]x[-180°, 180°] están permitidos [5]. Esto se debe a que hay varios átomos en las unidades i-1, i e i+1 cuyas distancias relativas dependen de (ϕ_i, ψ_i), de manera tal que algunos

valores de estos ángulos ocasionan que los átomos ocupen el mismo lugar en el espacio, lo cual es físicamente imposible (a esto se denomina repulsión estérica). Al eliminar todos los valores de (ϕ_i, ψ_i) que generan repulsiones estéricas entre aquellos átomos cuyas distancias relativas sólo dependen de (ϕ_i, ψ_i), se obtienen las regiones accesibles del plano (ϕ_i, ψ_i).

La rigidez del enlace C_i—N_{i+1} tiene como consecuencia la independencia de estas regiones entre unidades vecinas [5]. Esto quiere decir lo siguiente: los valores angulares permitidos para (ϕ_{i-1}, ψ_{i-1}) y (ϕ_i, ψ_i), definidos como aquellos ángulos que no producen repulsión estérica entre átomos cuyas distancias relativas dependen solamente de ϕ_{i-1}, ψ_{i-1}, ϕ_i y ψ_i, están formados por el producto cartesiano de las regiones admisibles obtenidas independientemente para las unidades i-1 e i.

Estas regiones locales "distendidas" pueden ser modeladas a través de un potencial local $U^R_i(\phi_i, \psi_i)$ que sólo depende de las coordenadas torsionales de la unidad i [5], llamado potencial local de Ramachandran, en honor a G. N. Ramachandran, quien fue el primer investigador que estudio estos efectos [14]. Las cuencas atractivas de U^R_i son llamadas precisamente cuencas atractivas de Ramachandran, y contienen todas las conformaciones de (ϕ_i, ψ_i) que no generan incompatibilidades geométricas locales.

El potencial U^R_i depende del tipo de aminoácido correspondiente a la unidad i. Se encuentra que existen solamente 4 tipologías de cuencas diferentes, a las cuales pertenecen los 20 aminoácidos. Estas tipologías son indicadas en la Figura 5, y vemos todas ellas están formadas por distintas combinaciones de las "mismas" cuencas. Llamaremos cuenca I a la contenida en el segundo cuadrante del plano (ϕ, ψ), cuenca II a la contenida en el tercer cuadrante, cuenca III a la contenida en el primer cuadrante y cuenca IV a la contenida en el

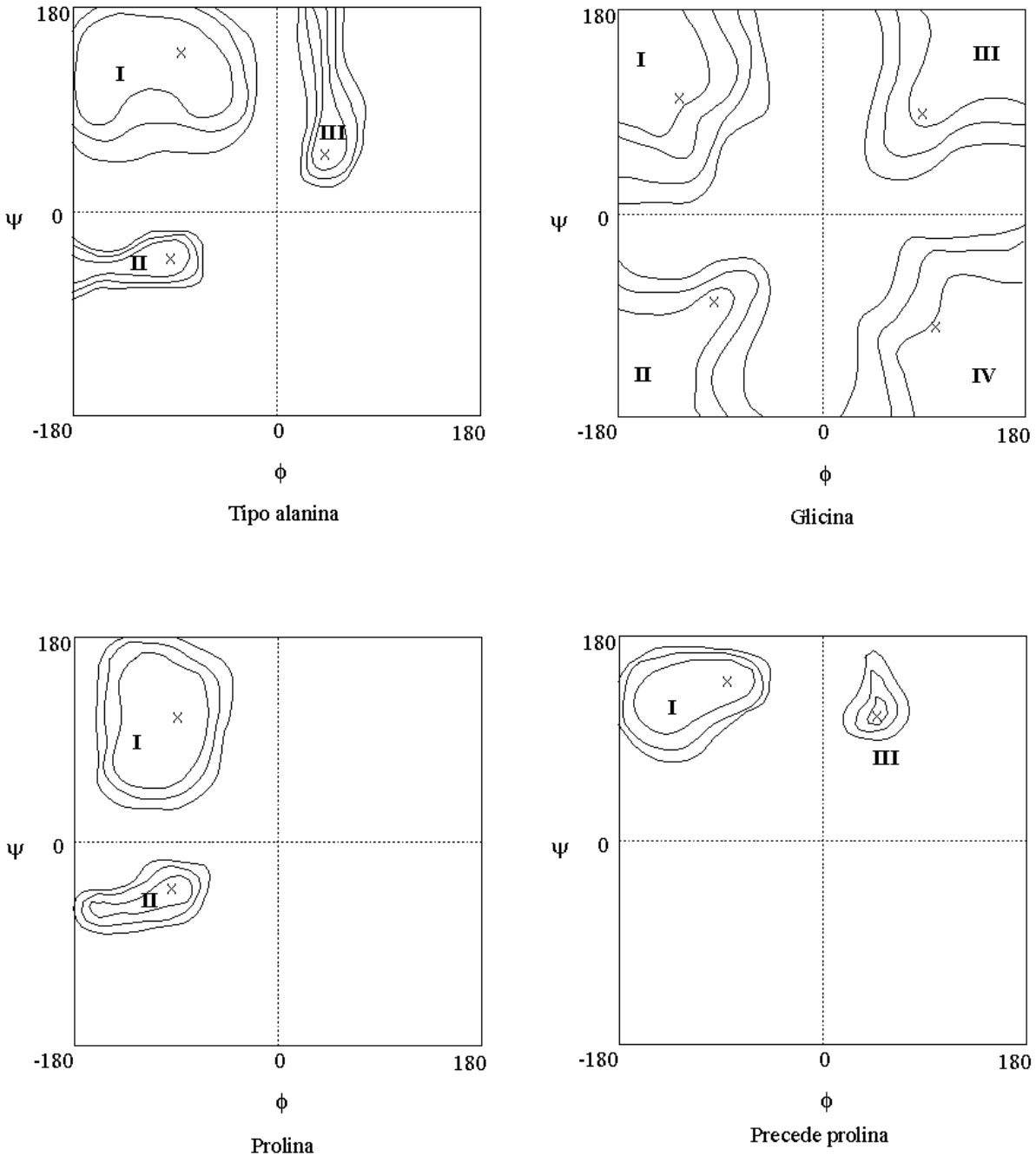


Figura 5: Las 4 tipologías de Ramachandran. A la tipología tipo alanina pertenecen todos los aminoácidos, excepto glicina, prolina y cualquier residuo que precede a la prolina. Los gráficos mostrados son representaciones esquemáticas de las curvas de nivel del potencial local de Ramachandran correspondiente, las cruces indican los mínimos locales en cada cuenca.

cuarto cuadrante. La glicina es el único aminoácido que tiene disponible las 4 cuencas, mientras que la tipología más común, llamada tipología alanina, carece de la cuenca IV. La tipología de la prolina consiste solamente en las cuencas I y II, mientras que cualquier aminoácido que precede a la prolina carece de las cuencas II y IV.

Generalmente la cuenca I es llamada cuenca "extendida" porque contiene los valores torsionales propios de las hojas β, mientras que las cuencas II y III son denominadas cuencas "compactas" debido a que contienen las coordenadas de las dos clases de hélices α: la dextrógira en la cuenca II y la levógira en la cuenca III. Finalmente, la cuenca IV, propia de la glicina, contiene coordenadas de β-turns.

La importancia de las cuencas de Ramachandran reside en dos hechos:

(a) En las estructuras nativas, los valores (ϕ, ψ) de la mayoría de las unidades permanecen dentro de las cuencas de Ramachandran [15]. Esto indica que las estructuras estables o mínimos del potencial global de una proteína están contenidos en una combinación o producto de cuencas de Ramachandran: para cada i las coordenadas torsionales i-ésimas del mínimo, (ϕ_i, ψ_i), están contenidas en alguna cuenca de Ramachandran de la unidad i.

(b) El tiempo que demanda una transición conformacional dentro de las cuencas de Ramachandran es mucho menor que el tiempo necesario para cambiar de una cuenca de Ramachandran a otra. Esto tiene como consecuencia que la dinámica global del sistema está "subordinada" a las dinámicas locales de Ramachandran: la proteína explora las posibilidades conformacionales que son accesibles dentro de una combinación o producto de cuencas determinado, antes de efectuar un cambio a otro producto de cuencas.

Estos elementos serán fundamentales para construir nuestro modelo del plegamiento en el próximo capítulo, en el cual utilizaremos una discretización o "digitalización" del espacio de conformaciones torsionales utilizando para ello a las cuencas de Ramachandran. Anticipándonos, diremos que la digitalización "módulo cuencas" es directamente y naturalmente sugerida por las restricciones geométricas del movimiento del objeto que nos ocupa.

1.6 El problema del plegamiento de proteínas

Luego de haber analizado en detenimiento las características físicas que determinan la estructura de una proteína y luego de haber visto las simplificaciones que es posible introducir en su descripción, estamos en condiciones de volver al problema del plegamiento.

La observación central en el cual se origina este problema es la siguiente: Se tiene un ensamble formado por una gran cantidad de proteínas naturales con la misma secuencia primaria, inmersas en un solvente que usualmente es agua. Todas las moléculas se hallan en alguna conformación desplegada carente de características estructurales definidas, llamada genéricamente ovillo aleatorio (random coil). En un lapso de tiempo que va desde los milisegundos a los pocos segundos, la mayor parte de las moléculas adoptan una estructura espacial bien determinada, en la cual permanecen. Este constituye el proceso de plegamiento y se observa que posee tres características esenciales: expeditividad, robustez y reproducibilidad. A continuación explicaremos cada una de estas características y sus consecuencias a la hora de construir un modelo realista del plegamiento:

(a) Expeditividad: La molécula de proteína es capaz de encontrar su estructura nativa de una manera muy veloz y eficaz, en medio de un enorme espacio conformacional y en tiempos incomparablemente más cortos que los tiempos ergódicos o termodinámicos. En efecto, hagamos

una estimación burda del tamaño del espacio de conformaciones de una proteína y del tiempo que demandaría una exploración exhaustiva del mismo. Supongamos que cada aminoácido puede adoptar 4 conformaciones torsionales distintas (lo cual es extremadamente conservador). Entonces, si la proteína tiene longitud N, existen 4^N conformaciones diferentes. Si la proteína fuera capaz de visitar 10^9 conformaciones por segundo, entonces, para N = 100, serían necesarios $4^{100}/10^9$ segundos para explorar todos las configuraciones posibles, y esto es aproximadamente 10^{40} años, un período de tiempo más largo que la edad del universo. Los experimentos muestran que la proteína encuentra su estructura nativa en el orden de segundos. Esto es lo que se conoce usualmente como paradoja de Levinthal [16]: el tiempo de plegamiento es inconmensurablemente menor que el requerido por una exploración exhaustiva del espacio de conformaciones.

La paradoja de Levinthal sería verdaderamente una paradoja si la proteína efectivamente alcanzara su estructura nativa a través de una búsqueda exhaustiva en el espacio de conformaciones. Lo que realmente indica esta discrepancia es que el proceso de plegamiento no ocurre en las escalas ergódicas o termodinámicas, sino que es un proceso puramente dinámico que ocurre muy lejos del equilibrio. Otro elemento experimental que justifica este hecho es que la estructura nativa no corresponde al mínimo global de la energía potencial, sino que es un mínimo alcanzado cinéticamente, muchas veces de naturaleza meta-estable [17].

Esto indica que los métodos para predecir estructuras tienen que ser necesariamente dinámicos o cinéticos, no termodinámicos, y deben ser capaces de generar los diferentes caminos de plegamiento individuales en las escalas de tiempo relevantes al problema.

(b) Robustez: Las moléculas de proteína recuperan la estructura nativa a pesar de que se apliquen cambios importantes en el ambiente de plegamiento (por ejemplo, modificando la temperatura o el pH del solvente), e incluso cambios en la misma secuencia primaria.

Otro elemento que también señala la robustez del proceso es la independencia con respecto al estado inicial, ya que cualquier conformación random coil conduce invariablemente a la misma estructura final. Si el plegamiento fuese un proceso termodinámico, este hecho no sería relevante, ya que la distribución de equilibrio es independiente del estado inicial, pero dado que el plegamiento es un proceso cinético, la independencia del estado inicial es señal también de la robustez del proceso.

Debido a su robustez, concluimos que dinámica de plegamiento no puede depender de los detalles conformacionales finos de la proteína, lo cual indica que es posible la adopción de una representación simplificada o "grosera" del sistema.

(c) Reproducibilidad: El plegamiento es altamente reproducible, ya que la estructura final generada es invariablemente la misma. Por otra parte, el proceso es de naturaleza estadística, y siempre hay un porcentaje de moléculas que no logran plegarse exitosamente. Esto indica que debemos recurrir a la teoría de los procesos estocásticos para poder describir adecuadamente el comportamiento promedio de un ensamble de moléculas y las excepciones que eventualmente puedan ocurrir durante el transcurso plegamiento.

En vista de estos tres elementos empíricos que caracterizan al plegamiento, y sus implicancias concretas en el plano teórico, los objetivos a alcanzar con esta tesis resultan ser los siguientes:

(a) Introducir una representación simplificada o "grosera" de una proteína que capture las características físicas que hacen del plegamiento un proceso expeditivo, robusto y reproducible.

(b) Construir un potencial efectivo en T^{2N} que represente las interacciones esenciales en una proteína y que incorpore al solvente de una manera implícita.

(c) Definir un proceso estocástico en el espacio de conformaciones "grosero" y probar que este proceso es la proyección rigurosa de la dinámica original del plegamiento.

(d) Diseñar e implementar un algoritmo que permita generar caminos individuales de plegamiento y que reproduzca el comportamiento estadístico de un ensamble de proteínas.

Es importante señalar que nuestros caminos de plegamiento serán generados completamente *ab-initio*, es decir que no se utilizará ninguna información referente a la estructura nativa. Esto parece una observación innecesaria, pero no lo es debido a que existen numerosos métodos para generar caminos de plegamiento que recurren de alguna manera a la estructura nativa. Por ejemplo, es muy común utilizar el llamado potencial de Gô [18], en el cual son penalizadas las conformaciones que durante el plegamiento se alejan de la estructura nativa. Por supuesto que estos métodos carecen de cualquier poder predictivo, y por el otro lado, asumen que el plegamiento es un proceso en el cual se avanza monotónicamente hacia la estructura nativa. Nuestras simulaciones muestran, por el contrario, que esta suposición no es cierta: las proteínas generalmente adoptan conformaciones imperfectas durante el plegamiento, que son necesarias para alcanzar finalmente la estructura nativa.

Con el propósito de verificar la utilidad práctica de nuestro modelo, estudiaremos una cuestión de gran importancia en el ámbito de la cinética del plegamiento: el llamado plegamiento de dos estados. Numerosos experimentalistas han reportado que una gran cantidad de proteínas naturales alcanzan su estructura nativa a través del plegamiento de dos estados [19, 20]: a lo largo del camino de plegamiento, sólo se observan dos estados con un peso estadístico apreciable: el segundo correspondiente a la estructura nativa final, mientras que el primer estado

corresponde en realidad a un ensamble de conformaciones estructuralmente relacionadas cuya obtención es precedida por una gran búsqueda conformacional, llamado ensamble del Estado de Transición (Transition State, TS). Una vez que este ensamble TS es alcanzado luego de un arduo período de "prueba y error", la molécula "cae" o "colapsa" fácilmente en la estructura nativa. Es por esto que el ensamble TS también es llamado núcleo colapsante, y se cree que contiene las características estructurales básicas de la estructura nativa, las cuales son designadas habitualmente con el nombre de topología nucleante. Se supone que [19-22]: **(a)** la obtención del núcleo colapsante depende de la ocurrencia concertada y cooperativa de varios eventos de formación estructural que involucran a algunos residuos clave y que conducen a la formación de la topología nucleante; **(b)** el proceso de colapso hacia la estructura nativa está constituido por el refinamiento progresivo de la topología nucleante.

Sin embargo, no existe todavía una concenso con respecto a la validez del modelo de dos estados. Todas las mediciones consisten en observar señales que están relacionadas indirectamente los eventos del plegamiento que ocurren en la realidad, y en consecuencia existen interpretaciones diferentes de las mismas observaciones. Estas dificultades en el aspecto experimental se deben a que: **(a)** las proteínas son objetos muy pequeños, de dimensiones moleculares, **(b)** los eventos de plegamiento ocurren en escalas de tiempo extremadamente breves, muy inferiores al segundo, y **(c)** no existe un teoría que brinde un marco conceptual adecuado para los experimentos [22-24]. Por lo tanto, un modelo teórico que permita reproducir la dinámica del plegamiento, en especial en los rangos de tiempo asociados con la formación del núcleo colapsante, sería de gran utilidad para resolver estas cuestiones.

Hemos seleccionado dos proteínas que están típicamente vinculadas con el modelo de dos estados: ubiquitina (1ubi) [22, 23] y la variante hipertermófila de la proteína G (1gb4) [24], las cuales estudiaremos en detalle con nuestro algoritmo de plegamiento, en lo que se refiere a:

(a) La comprobación computacional del modelo de dos estados: ¿Existe realmente un ensamble TS para 1ubi y 1gb4 a partir del cual el plegamiento colapsa hacia la estructura nativa?

(b) El análisis del ensamble TS: ¿Qué conformaciones forman el ensamble TS? ¿Cuál es la relación estructural entre las mismas? ¿Por qué la búsqueda conformacional disminuye una vez que el ensamble TS es alcanzado?

(c) El estudio de la diversidad de caminos y la cooperatividad en el plegamiento: ¿Existen varios caminos posibles para llegar al ensamble TS? ¿Cuáles son los residuos clave cuyas interacciones cooperativas posibilitan la formación del ensamble TS? ¿Es posible efectuar modificaciones en las secuencias primarias (mutaciones puntuales) de 1ubi y 1gb4 para aumentar los efectos cooperativos y disminuir el "cuello de botella" cinético asociado con la formación del TS?

1.7 Dinámica Molecular: un primer intento de resolver el problema del plegamiento

En esta sección mostraremos con ejemplos concretos la gran limitación de los métodos "directos" o "clásicos" para resolver el problema del plegamiento. Como ya se dijo antes, la manera más inmediata de tratar matemáticamente el problema es a través de las ecuaciones de Newton en $R^{N_p+N_s}$, utilizando para ello el potencial riguroso U que incorpora todas las interacciones de nivel atómico. Estos métodos son designados genéricamente con el nombre de Dinámica Molecular con todos los átomos (all-atom Molecular Dynamics) [6].

Sea **x**(t) el vector posición del sistema en el instante t, de manera tal que ($x_{3(i-1)+1}$(t), $x_{3(i-1)+2}$(t), $x_{3(i-1)+3}$(t)) son las coordenadas del átomo i-ésimo en el espacio tridimensional. Sea **M** la matriz que contiene las masas de cada uno de los átomos del sistema: $M_{ij} = m_i$ si $3_{(i-1)+1} \leq j \leq 3_{(i-1)+3}$, 0 en caso contrario, y con m_i = masa del átomo i-ésimo. Entonces la 2da Ley de Newton puede ser escrita como:

$$\mathbf{M}\, d^2\mathbf{x}(t)/dt^2 = -\text{grad}\, U(\mathbf{x}(t)) \qquad (1)$$

donde las condiciones iniciales **x**(0) y **v**(0) deben seleccionarse de manera tal que correspondan a una estructura random coil. El problema consiste en hallar la solución **x**(t) de esta ecuación para t en el intervalo [0, t_{total}], donde t_{total} pertenece a los rangos de interés biológicos (de 1 milisegundo a algunos segundos).

El paso de tiempo que hay que utilizar para resolver numéricamente la ecuación (1) está determinado por los modos de movimiento más rápidos del sistema [6], y éstos resultan ser del orden de 1-5 fs (1 fs = 10^{-15} s), correspondientes a los períodos típicos de las vibraciones atómicas. Este hecho, sumado a que el número de coordenadas es enorme y a que el potencial U es muy complejo, hace que los métodos de Dinámica Molecular sean prácticamente inalcanzables en términos computacionales. Consideremos dos ejemplos al respecto:

(a) La primera simulación de dinámica molecular que logró acercarse mínimamente a las escalas biológicas de interés fue realizada en 1998 sobre un pequeño fragmento de 36 aminoácidos [25]. Se utilizó una supercomputadora Cray T3C con 256 procesadores y requirió 6 meses de tiempo de cómputo para generar un solo camino de 1 μs (1 μs = 10^{-6} s) en tiempo real.

(b) IBM se encuentra diseñando en este momento una supercomputadora masivamente paralela cuyo único objetivo será efectuar simulaciones de dinámica molecular (proyecto Blue Gene)

[26]. La misma contará con aproximadamente 50000 procesadores y necesitará un año completo de cómputo para simular el plegamiento de una proteína de 300 aminoácidos, de acuerdo con cálculos de la misma IBM.

Finalmente hay que señalar que una sola simulación de dinámica molecular es insuficiente para construir un modelo que represente el comportamiento estadístico de un ensamble de proteínas.

1.8 Referencias

[1] *Molecular Cell Biology.* J. Darnell, H. Lodish, D. Baltimore. W. H. Freeman and Company (1986).

[2] Biocatalisis: synthesis methods that exploit enzymatic activities. Editores: P. Ball, K. Z. Ziemelis. *Nature* **409**, 225-268 (2001).

[3] Aging. Editores: B. Pulverer, R. Turner, R. Drand. *Nature* **408**, 231-269 (2000).

[4] Cancer. Editores: B. Pulverer, L. Anson, C. Surridge. *Nature* **411**, 335-395 (2001).

[5] *Biophysical Chemistry, Part I: The Conformation of Biological Macromolecules*. C. R. Cantor, P. R. Schimmel. W. H. Freeman and Company (1980).

[6] *Proteins: a Theoretical Perspective of Dynamics, Structure and Thermodynamics*. C. Brooks III, M. Karplus, B. M. Pettit. Advances in Chemical Physics, Volumen LXXI. John Wiley and Sons (1988).

[7] Effective energy function for proteins in solution. T. Lazaridis, M. Karplus. *Proteins: Struct. Funct. Genet.* **35**, 132-152 (1999).

[8] Free energies of hydration of solute molecules 3. Application of the hydration shell model to charged organic molecules. Y. K. Kang, G. N'emethy, H. A. Scheraga, H. A. *J. Phys. Chem.* **91**, 4118-4123 (1987).

[9] Conformational-dependent environments in folding proteins. A. Fernández. *J. Chem. Phys.* **114**, 2489-2502 (2001).

[10] Self-organization and mismatch tolerance in protein folding. General theory and an application. A. Fernández, R. S. Berry. *J. Chem. Phys.* **112**, 5212-5222 (2000).

[11] Cooperative walks in a cubic lattice: Protein folding as a many-body problem. A. Fernández. *J. Chem. Phys.* **115**, 7293-7297 (2001).

[12] Areas, volumes, packing and protein structure. F. M. Richards. *Ann. Rev. of Biophys. Bioeng.* **6**, 151-176 (1977).

[13] Discrimination of the native from misfolded protein models with an energy function including implicit solvation. T. Lazaridis, M. Karplus. *J. Mol. Biol.* **288**, 477-487 (1998).

[14] Stereochemistry of polypeptide chain configurations. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan. *J. Mol. Biol.* **7**, 95-98 (1963).

[15] Protein Structures: The End Point of the Folding Pathway. J. M. Thornton. En: *Protein Folding*. Editor: T. A. Creighton. W. H. Freeman and Company (1992).

[16] Levinthal's paradox. R. Zwanzig, A. Szabo, B. Bagchi. *Proc. Natl. Acad. Sci. USA* **89**, 20-22 (1992).

[17] Is there a code for protein folding? R. Jaenicke. En: *Protein Structure and Protein Engineering*. Editores: E. L. Winnaker, R. Huber. Springer (1988).

[18] Studies in protein folding, unfolding and fluctuations by computer simulation. H. Taketomi, Y. Ueda, N. Gô. *Int. J. Pept. Protein Res.* **7**, 445-459 (1975).

[19] The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the protein folding nucleus. Z. Guo, D. Thirumalai. *Fold. Des.* **2**, 377-391 (1997).

[20] Specific nucleus as the transition state for protein folding: evidence from the lattice model. V. I. Abkevich, A. M. Gutin, E. I. Shakhnovich. *Biochemistry* **33**, 10026-10036 (1994).

[21] Effects of point mutations on the folding of globular proteins. C. R. Matthews. *Methods Enzymol.* **154**, 498-511 (1987).

[22] Distinguishing between two-state and three-state models for ubiquitin folding. B. A. Krantz, T. R. Sosnick. *Biochemistry* **39**, 11696-11701 (2000).

[23] Evidence for a 3-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. S. Khorasanizadeh, I. D. Peters, H. Roder. *Nature Str. Biol.* **3**, 193-205 (1996).

[24] Design, structure and stability of a hyperthermophile protein variant. S. M. Malakauskas, S. L. Mayo. *Nature Str. Biol.* **5**, 470-475 (1998).

[25] Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Y. Duan, P. A. Kollman. *Science* **282**, 740-744 (1998).

[26] Página web del proyecto IBM Blue Gene: http://www.research.ibm.com/bluegene

Capítulo 2

Un nuevo modelo matemático para el plegamiento:

¿Cuándo es admisible simplificar?

2.1 Dinámica de plegamiento en T^{2N}

En el capítulo anterior concluimos que es posible describir la conformación de una proteína y su dinámica de plegamiento en el espacio de estados torsionales del backbone: $X = T^{2N}$. A lo largo de este capítulo asumiremos que existe una función U^{eff}: $X \rightarrow R$ que representa la energía potencial como función de las coordenadas torsionales de la proteína.

Ya que X es una variedad diferenciable, las ecuaciones del movimiento en X pueden ser planteadas de manera natural utilizando el formalismo propio de la geometría diferencial [1]. Sean TX y TTX los espacios tangente y segundo tangente de X (los espacios de velocidades y aceleraciones sobre X, respectivamente). Sea F: $TX \rightarrow TTX$ el campo vectorial sobre TX determinado por el gradiente de V: $TX \rightarrow R$, con $V(v_x) = U^{eff}(x)$, es decir F = -grad V. El campo F define las ecuaciones de movimiento sobre X. Una curva α: $[t, t_{total}] \rightarrow TX$ se dice una curva integral del sistema si cumple:

$$\alpha'(t) = F(\alpha(t)) \qquad (1)$$

Si α es una curva integral en TX y τ_X: $TX \rightarrow X$ denota la proyección canónica del espacio tangente sobre X, es decir, $\tau_X(v_x) = x$, entonces la curva integral en X está dada por:

$$\beta(t) = \tau_X(\alpha(t))$$

Esta constituye la formulación lagrangiana del problema. Si escribimos la ecuación (1) en coordenadas, obtendremos precisamente las ecuaciones de Euler-Lagrange [1], las cuales determinan un sistema de ecuaciones diferenciales de segundo orden. Una formulación equivalente del problema consiste en construir la función hamiltoniana del sistema y plantear las ecuaciones de Hamilton en el espacio cotangente T*X, que representa el espacio de fases del sistema, es decir, el espacio de configuraciones y momentos [1].

Sin embargo, una curva solución de (1) representa la trayectoria de una sola molécula, y como ya se dijo en el capítulo anterior, estamos interesados en obtener una descripción estadística de un ensamble de moléculas que comienza su evolución desde una configuración random coil cualquiera. Esta descripción estadística queda determinada por la función de densidad $\rho(x, t)$, donde $x \in X$ y $t \geq 0$. La probabilidad de encontrar a una molécula del ensamble en la región $R \subseteq X$ en el instante t es:

$$P(R, t) = \int_R \rho(x, t)\, dx$$

La densidad inicial $\rho(x, 0) = \rho_0(x)$ corresponde a una distribución uniforme en la región de X que contiene a las conformaciones random coil. El problema es entonces encontrar la densidad $\rho(x, t)$ que cumpla con esta condición inicial y que represente el comportamiento del ensamble bajo los efectos del potencial U^{eff}.

Este problema se vuelve matemáticamente más manejable si el proceso estocástico en X que determina la densidad $\rho(x, t)$ es markoviano y homogéneo en el tiempo (estacionario) [2-4]. La hipótesis markoviana es válida [2] si las escalas de tiempo que deseamos observar con nuestro proceso estocástico son mayores que los tiempos de termalización del sistema [7]. Dado

que la termalización ocurre mucho más rápidamente que las transiciones torsionales, podemos aceptar de aquí en adelante la markovianidad del proceso de plegamiento en X.

Bajo estas hipótesis, la ecuación maestra que nos permite obtener $\rho(x, t)$ es [2]:

$$\partial\rho(x, t)/\partial t = \int_X \{W(x, x')\rho(x', t) - W(x', x)\rho(x, t)\}\, dx' \qquad (2)$$

con condición de borde $\rho(x, t) = \rho_0(x)$ y donde $W(x, x')$ son las velocidades instantáneas de transición [2].

La ecuación maestra en T^{2N} ha sido ampliamente estudiada, tanto desde el punto de vista teórico [5], así como también desde el punto de vista numérico [6]. La importancia de la ecuación maestra reside en el hecho de que provee el marco formal con el se caracteriza estadísticamente la dinámica torsional del plegamiento. El estudio teórico de esta ecuación ha permitido deducir rigurosamente propiedades importantes acerca del proceso de plegamiento [5, 6]. Sin embargo, la aplicabilidad de la ecuación maestra a efectos de generar trayectorias de plegamiento es más limitada. La obtención de $\rho(x, t)$ requiere resolver una ecuación diferencial de dimensión 2N, lo cual puede resultar ser una tarea casi imposible, especialmente para proteínas grandes. Aún en el caso de haber obtenido la función de densidad $\rho(x, t)$, el problema todavía no está resuelto: por ejemplo, si deseamos obtener la trayectoria promedio $\langle x(t)\rangle$, necesitamos evaluar la siguiente integral: $\langle x(t)\rangle = \int_X x\, \rho(x, t)\, dx$, lo cual está fuera de la capacidad computacional actual [6].

Estas dificultades son justamente las que motivan nuestro modelo discretizado del plegamiento, el cual será presentado y discutido en el capítulo siguiente. Sin embargo, la ecuación maestra torsional (2) constituye el punto de partida que utilizaremos para generar una dinámica discreta del plegamiento.

2.2 Discretización de T^{2N}

Hemos visto hasta ahora que todos los modelos matemáticos que intentan dar una descripción rigurosa del proceso de plegamiento tienen limitaciones importantes en el momento de ofrecer una solución concreta del problema, dado que las ecuaciones que proveen no son tratables numéricamente, al menos en las escalas de tiempo en las cuales ocurren los eventos centrales del plegamiento. Este es el motivo práctico que nos lleva a introducir un nuevo paso de reducción o simplificación representacional con el propósito de hacer computacionalmente accesible al problema del plegamiento. Recordemos que desde el punto de vista teórico, una representación simplificada representaría además una racionalización de la robustez del proceso de plegamiento. Por lo tanto, tenemos suficientes razones, tanto prácticas como teóricas, para efectuar una nueva simplificación conformacional.

La subordinación del potencial global a los movimientos locales entre cuencas de Ramachandran (ver sección 1.5 del capítulo anterior) sugiere que es posible reproducir la dinámica del plegamiento en una descripción discreta en la cual el estado la molécula queda caracterizado al indicar solamente la cuenca de Ramachandran que ocupa cada unidad de la cadena, en vez de indicar sus coordenadas intrínsecas específicas. En términos concretos, el problema a tratar en este capítulo puede formularse con la pregunta: ¿Cuándo puede justificarse como válida la simplificación de la dinámica torsional de una proteína que resulta de pasar al cociente definiendo una relación de equivalencia "módulo cuencas de Ramachandran"?

Antes de seguir adelante, fijemos la siguiente notación:

(a) $X_i = T^2$

(b) $x_i = (\phi_i, \psi_i)$

(c) $X = \prod_{i=1...N} X_i$

De esta manera, una conformación torsional de la cadena puede ser escrita como $x = (x_1, x_2,...x_N)$, con $x_i = (\phi_i, \psi_i) \in X_i$. A efectos de llegar a una discretización de X, empezaremos por definir una partición de cada espacio torsional local X_i utilizando para ello a las cuencas de Ramachandran del residuo i:

Definición 1: $x_i \sim_i x'_i$ si y sólo si x_i y x'_i pertenecen a la misma cuenca de atractiva del potencial U_i^R. Notaremos con Y_i al conjunto de clases de equivalencia $X_i/\sim$, y π_i: $X_i \rightarrow Y_i$ a la proyección canónica: $\pi_i(x_i)$ = clase de equivalencia de x_i.

Y_i está formado a lo sumo por 4 elementos distintos, ya que hay como máximo 4 cuencas de Ramachandran (ver sección 1.5 del capítulo anterior). En esta definición aparece un problema técnico que es el siguiente: las separatrices de U_i^R no pertenecen a ninguna cuenca atractiva. Sin embargo el conjunto de separatrices tiene medida nula en X_i , por lo cual los cálculos y resultados que presentaremos más adelante serán aplicables de todas maneras, ya que no se ven afectados por "modificaciones de medida nula" en las particiones.

En base a esta relación definida para cada coordenada x_i podemos introducir de manera natural una partición en todo el espacio X:

Definición 2: $x \sim x'$ si y sólo sí $x_i \sim_i x'_i$ para todo i. Notaremos con Y al conjunto de clases de equivalencia $X/\sim$, y π: $X \rightarrow Y$ a la proyección canónica: $\pi(x)$ = clase de equivalencia de x.

De estas definiciones resulta que:

$$Y = \prod_{i=1...N} Y_i$$

$$\pi(x) = \pi(x_1, x_2,...x_N) = (\pi_1(x_1), \pi_2(x_2) \ldots \pi_N(x_N))$$

Podemos visualizar a $y \in Y$ como un "producto de cuencas locales de Ramachandran": $y = (y_1, y_2 \ldots y_N)$, $y_i \in Y_i$. Es decir, y es un vector formado por una asignación de cuencas de Ramachandran para cada unidad de la cadena. Llamaremos LTM (Local Topology Matrix, Matriz de Topología Local) a estos vectores.

El número de elementos de Y (el número de LTM's distintas) es $R = R_1R_2...R_N$, donde R_i es el número de cuencas atractivas de U_i^R. Como R_i es en promedio igual a 3, Y tiene aproximadamente 3^N elementos.

Nuestro problema es entonces construir una dinámica markoviana en $Y = \pi(X)$ que sea "compatible" con la dinámica rigurosa en X. En principio, tenemos dos maneras para generar dinámicas markovianas en Y: una es resolviendo la ecuación maestra (2) para obtener $\rho(x, t)$ y luego proyectar esta densidad sobre Y de la manera natural: $\pi(\rho)(y, t) = \int_y \rho(x, t)\, dx$. Por supuesto que no ganamos demasiado con este método, ya que requiere resolver (2). Una segunda manera es definir una cadena de Markov en Y especificando sus velocidades de transición, y luego resolver la ecuación maestra correspondiente en Y para obtener P(y, t). La condición que debe cumplirse para que la cadena en Y sea compatible con el proceso original en X debe ser la siguiente: "$P = \pi(\rho)$". Veremos que es posible construir una cadena en Y que cumpla esta condición, siempre y cuando la cadena X sea "consistente" con la partición Y. Desde el punto de vista matemático, estos resultados son, hasta donde sabemos, totalmente nuevos, ya que hasta el momento no se han estudiado proyecciones de procesos estocásticos con el propósito de

describir una dinámica discretizada que constituye un granulado grueso de la dinámica detallada original. Remarquemos nuevamente que este sacrificio en resolución estructural es absolutamente necesario para abarcar computacionalmente los tiempos de plegamiento de un objeto tan complejo como lo es una proteína natural. Vale decir que la inspiración para nuestro tratamiento procede de razones prácticas y no puramente matemáticas.

Seguiremos la notación utilizada en [2], ya que representa un buen compromiso entre rigor matemático e interpretación física, balance de gran importancia en un trabajo de matemática aplicada como este. Tengamos presente que muchos autores en el campo de los procesos estocásticos [4] admiten que la teoría general de procesos estocásticos, a pesar de su gran desarrollo matemático, es muy difícil de aplicar en casos concretos debido a su gran complejidad notacional y técnica.

Los resultados que obtendremos en las secciones 2.4 y 2.5 son generales ya que el espacio X y su partición Y son arbitrarios, excepto por la condición de que en X podamos calcular integrales, es decir, que X tenga definido un diferencial de volumen dx.

2.3 Algunos resultados preliminares acerca de procesos estocásticos

El propósito de esta sección es introducir la notación y enunciar algunos resultados fundamentales acerca de procesos estocásticos. Para mayores detalles, consultar [2], [3] o cualquier otro libro introductorio de procesos estocásticos.

Consideremos un proceso estocástico definido en un espacio X. El mismo queda caracterizado totalmente al dar las funciones de densidad siguientes para todo n y para todo $x_i \in X$ y todo $t_i \geq 0$:

$$\rho_n(x_n, t_n; x_{n-1}, t_{n-1};...x_1, t_1)$$

La probabilidad de que la conformación del sistema pertenezca a la región $X_n \subseteq X$ en el instante t_n, a $X_{n-1} \subseteq X$ en el instante t_{n-1}, etc., es:

$$P_n\,(X_n, t_n;\, X_{n-1}, t_{n-1};...X_1, t_1) = \int_{X_1 \ldots X_n} \rho_n(x_n, t_n;\, x_{n-1}, t_{n-1};...x_1, t_1)\, dx_n...dx_1$$

Se dice que el proceso es markoviano si las densidades condicionales, cuando están definidas, verifican:

$$\rho_{1|n}(x_{n+1}, t_{n+1}|x_n, t_n;...x_1, t_1) = \rho_{1|n}(x_{n+1}, t_{n+1}|x_n, t_n)$$

El proceso es homogéneo o estacionario cuando las densidades condicionales $\rho_{1|1}$ no dependen del instante inicial sino que sólo dependen del tamaño del intervalo:

$$\rho_{1|1}(x, t|x', t') = \rho_{1|1}(x, t-t'|x', 0)$$

En estas circunstancias (proceso markoviano y homogéneo) definimos a la densidad de transición entre x' y x en un intervalo t como:

$$T_t(x|x') = \rho_{1|1}(x, t|x', 0)$$

Estas funciones de densidad verifican las ecuaciones de Chapman-Kolmogorov:

$$T_{t+t'}(x|x') = \int_X T_t\,(x|x'')T_{t'}(x''|x')\, dx'' \qquad (3)$$

Notemos que:

$$\rho_1(x, t) = \int_X \rho_1(x, t, x', 0)\, dx' = \int_X \rho_{1|1}(x, t|x', 0)\rho_1(x', 0)\, dx' = \int_X T_t(x|x')\rho_1(x', 0)\, dx' \qquad (4)$$

Por otro lado, la función de densidad ρ_n puede escribirse como:

$$\rho_n(x_n, t_n;\, x_{n-1}, t_{n-1};...x_1, t_1) = \rho_n(x_n, t_n|x_{n-1}, t_{n-1};...x_1, t_1)\rho_{n-1}(x_{n-1}, t_{n-1};...x_1, t_1) =$$

$$= T_{\Delta t_n}(x_n|x_{n-1})T_{\Delta t_{n-1}}(x_n|x_{n-1})...\ T_{\Delta t_2}(x_2|x_1)\rho_1(x_1, t_1) \qquad (5)$$

en donde $\Delta t_k = t_k - t_{k-1}$.

Recíprocamente, un conjunto de densidades de transición $T_t(x|x')$ que verifiquen las ecuaciones de Chapman-Kolmogorov (3), junto con una densidad inicial $\rho_1(x, 0) = \rho_0(x)$, definen completamente un proceso markoviano y homogéneo a través de (4) y (5).

Supondremos que el proceso markoviano definido por las densidades $T_t(x|x')$ es suficientemente regular de manera que existan los siguientes límites:

$$W(x, x') = \lim_{t \to 0^+} \{T_t(x|x')-T_0(x|x')\}/t$$

Las funciones W(x, x') son llamadas velocidades instantáneas de transición, y bastan para caracterizar totalmente al proceso, ya que es posible recuperar las densidades de transición a partir de ellas. Generalmente es más sencillo definir el proceso en término de las velocidades W(x, x'), ya que las mismas tiene una interpretación física muy clara [2].

La densidad $\rho_1(x, t)$, que a partir de ahora notaremos simplemente con $\rho(x, t)$, se obtiene a partir de la ecuación maestra que involucra a las velocidades de transición:

$$\partial\rho(x, t)/\partial t = \int_X \{W(x, x')\rho(x', t) - W(x', x)\rho(x, t)\}\, dx'$$

La condición de borde de esta ecuación lineal a derivadas parciales de primer orden está definida por $\rho(x, 0) = \rho_0(x)$, siendo ρ_0 la densidad inicial.

Si X es un conjunto finito, la ecuación maestra toma la siguiente forma:

$$dP_x(t)/dt = \sum_{x'} \{W_{xx'}P_{x'}(t)-W_{x'x}P_x(t)\}$$

Aquí $P_x(t)$ ya no es una densidad, sino que es directamente la probabilidad de que el sistema se encuentre en el estado x en el instante t. Vale que $W_{xx} = -\sum_{x'}W_{x'x}$. Utilizando esta igualdad podemos reescribir a la ecuación maestra como:

$$dP_x(t)/dt = \sum_{x'} W_{xx'}P_{x'}(t)$$

En notación matricial:

$$dP(t)/dt = WP(t) \qquad (6)$$

donde $W = (W_{xx'})$ y $P(t) = (P_x(t))$. La solución formal de (6) está dada por la función "exponencial de matrices":

$$P(t) = P(0)\, e^{Wt} = P(0) \sum_{n=0..\infty} W^n/n!\; t^n$$

2.4 Proyección de procesos estocásticos: el paso al cociente

Supongamos ahora que en el espacio X definimos una partición en clases disjuntas. Notaremos con Y al conjunto de las clases, y con π a la proyección canónica π: X $\rightarrow$ Y. Supondremos que Y es numerable (o finito). Notaremos con "x" a los elementos de X y con "y" a los elementos de Y, es decir, a las clases de la partición.

Si tenemos un proceso estocástico en X, la definición natural de proceso proyectado en Y es la siguiente:

Definición 3: Sean $\rho_n(x_n, t_n; x_{n-1}, t_{n-1};...x_1, t_1)$ las funciones de densidad que definen un proceso estocástico en X. Si Y es una partición de X, el proceso proyectado en Y está definido por las siguientes probabilidades:

$$P_n(y_n, t_n; y_{n-1}, t_{n-1};...y_1, t_1) = \int_{y_1...y_n} \rho_n(x_n, t_n; x_{n-1}, t_{n-1};...x_1, t_1)\, dx_n...dx_1$$

A continuación veremos que si el proceso en X es markoviano y homogéneo, su proyección no necesariamente lo es. La idea es construir una cadena en X donde haya caminos con probabilidad nula, tal como se muestra en la Figura 6: la probabilidad de ir de la región y_0 a y_1 - y'_1 es nula, por lo tanto $P_{1|2}(y_2, t_2|y_1, t_1; y_0, t_0) = P_{1|1}(y_2, t_2|y'_1, t_1) = p'_1$. Por otro lado,

$P_{1|1}(y_2, t_2|y_1, t_1) = P_{1|1}(y_2, t_2|y'_1, t_1)P_{1|1}(y'_1, t_1|y_1, t_1) + P_{1|1}(y_2, t_2|y_1 - y'_1, t_1)P_{1|1}(y_1-y'_1, t_1|y_1, t_1)$. Llamando p a $P_{1|1}(y'_1, t_1|y_1, t_1)$ podemos escribir $P_{1|1}(y_2, t_2|y_1, t_1) = p'_1p + p_1(1 - p)$. Supongamos que $p_1 = 3p'_1$. Entonces $P_{1|1}(y_2, t_2|y_1, t_1) = p'_1(3 - 2p)$. Para que $P_{1|2}(y_2, t_2|y_1, t_1; y_0, t_0) \neq P_{1|1}(y_2, t_2|y_1, t_1)$ bastaría que $p < 1$. Veamos como hacer esto en un caso concreto.

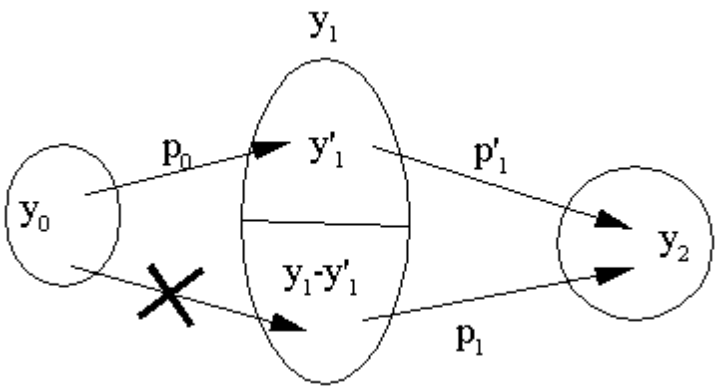


Figura 6: Esquema en el espacio de conformaciones X de una posible partición para la cual la proyección del proceso estocástico en X no es markoviana.

Ejemplo 1:

Consideremos un cadena de Markov homogénea de tiempo discreto (t = 0, 1, 2,...) sobre el espacio numerable X = {a, b, c, d,...}. Se tienen las siguientes probabilidades de transición de un paso:

$T_1(b|a) = 1$

$T_1(c|a) = 0$

$T_1(d|b) = 1/6$

$T_1(d|c) = 1/2$

Particionemos a X agrupando b y c, y dejando a los demás elementos en clases individuales. Entonces:

$$P_{1|2}(d, 2|\{b, c\}, 1; a, 0) = P_{1|1}(d, 2|b, 1) = T_1(d|b) = 1/6$$

$$P_{1|1}(d, 2|\{b, c\}, 1) = P_{1|1}(d, 2|b, 1)P_{1|1}(b, 1|\{b, c\}, 1) +$$

$$P_{1|1}(d, 2|c, 1)P_{1|1}(c, 1|\{b, c\}, 1)$$

$$= 1/6\ p + 1/2\ (1 - p) = (3 - 2p)/6$$

donde $p = P_{1|1}(b, 1|\{b, c\}, 1)$ y donde se utilizó que:

$$P(A|B \cup C) = P(A|B)P(B|B \cup C) + P(A|C)P(C|B \cup C)$$

cuando B y C son eventos disjuntos. Podemos construir una distribución inicial $P_1(x, 0)$ tal que $p < 1$. Observemos que si $0 < P_1(c, 1)$ entonces:

$$p = P_{1|1}(b, 1|\{b, c\}, 1) = P_1(b, 1)/\{P_1(b, 1) + P_1(c, 1)\} < 1$$

Esto se puede lograr, por ejemplo, si $T_1(c|c) > 0$ y $P_1(c, 0) > 0$.

La condición de markovianidad falla en este caso porque del estado 1 no se puede ir al 3, por lo tanto la condición {{b, c}, 1; a, 0} es equivalente a decir que el sistema está ocupando el estado 2 en el instante 1. Esto sugiere que para conservar la markovianidad necesitamos que no haya caminos con probabilidad nula.

Este mismo ejemplo sirve para mostrar que un proceso proyectado no es necesariamente homogéneo, a pesar de que el proceso de base lo sea. Hagamos el siguiente cálculo:

$$P_{1|1}(d, 1|\{b, c\}, 0) = P_{1|1}(d, 1|b, 0)P_{1|1}(b, 0|\{b, c\}, 0) +$$

$$P_{1|1}(d, 1|c, 0)P_{1|1}(c, 0|\{b, c\}, 0)$$

$$= 1/6\ q + 1/2\ (1 - q) = (3 - 2q)/6$$

donde $q = P_{1|1}(b, 0|\{b, c\}, 0)$. Ya vimos que $P_{1|1}(d, 2|\{b, c\}, 1) = (3 - 2p)/6$, por lo tanto , si $p \neq q$ resultará que $P_{1|1}(d, 2|\{b, c\}, 1) \neq P_{1|1}(d, 1|\{b, c\}, 0)$, es decir, el proceso proyectado no es homogéneo en el tiempo. Tomando $P_1(b, 0) = 0$, resulta que $q = 0$, y esta condición no impide tomar simultáneamente $p < 1$.

En este caso, se ve que la pérdida de homogeneidad temporal está relacionada con el hecho de que $P_{1|1}(b, 0|\{b, c\}, 0) \neq P_{1|1}(b, 1|\{b, c\}, 1)$. Esto sugiere que si las probabilidades condicionales dentro de las clases son constantes en el tiempo, sería posible que la proyección conserve la homogeneidad.

A continuación enunciaremos condiciones suficientes para que el proceso proyectado sea markoviano y homogéneo, basados en las observaciones hechas en el ejemplo anterior. En primer lugar, supondremos que cualquier camino en Y tiene probabilidad no nula, es decir, aceptaremos la siguiente condición:

(P0) $P_n(y_n, t_n; y_{n-1}, t_{n-1};...y_1, t_1) > 0$ para todo n, $y_n,...y_1 \in Y$ e instantes $t_n,...t_1 \geq 0$.

Observemos que esta condición se cumple, en particular, si las densidades ρ_n son estrictamente positivas.

Introduzcamos la siguiente definición, que nos permitirá simplificar los razonamientos subsiguientes:

Definición 4: La función $\sigma_{1|n}$ se define para todo n, $x_{n+1} \in X$, $y_n,...y_1 \in Y$ e instantes $t_n,...t_1 \geq 0$ como sigue:

$$\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \{\int_{y_1...y_n} \rho_{n+1}(x_{n+1}, t_{n+1},...x_1, t_1)\, dx_n...dx_1\}/P(y_n, t_n;...y_1, t_1)$$

$\sigma_{1|n}$ puede interpretarse como la densidad de la probabilidad de adoptar una conformación en X dado que se visitaron las clases y_n, ..., y_1 en los instantes t_n, ..., t_1.

Lema 1: $P_{1|n}(y_{n+1}, t_n|y_n, t_n;...y_1, t_1) = \int_{y_{n+1}} \sigma_{1|n+1}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1)\, dx_{n+1}$

Demostración:

$$P_{1|n}(y_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = P_{n+1}(y_{n+1}, t_{n+1;}\, y_n, t_n;...y_1, t_1)/P_n(y_n, t_n,...x_1, t_1)$$

$$= \{\int_{y_1...y_{n+1}} \rho_{n+1}(x_{n+1}, t_{n+1;}\, x_n, t_n;...x_1, t_1)dx_{n+1}...dx_1\}/P_n(y_n, t_n,...x_1, t_1)$$

$$= \int_{y_{n+1}} \{\int_{y_1...y_n} \rho_{n+1}(x_{n+1}, t_{n+1;}\, x_n, t_n;...x_1, t_1)dx_n...dx_1/P_n(y_n, t_n,...x_1, t_1)\}dx_{n+1}$$

Lema 2: Si $n > 1$, vale que:

$$\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \int_{y_n} \rho_{1|1}(x_{n+1}, t_{n+1}|x_n, t_n)\{\sigma_{1|n-1}(x_n, t_n|y_{n-1},...t_1)/P_{1|n-1}(y_n, t_n|y_{n-1}...t_1)\}\, dx_n$$

Si $n = 1$, entonces:

$$\sigma_{1|1}(x_2, t_2|y_1, t_1) = \int_{y_1} \rho_{1|1}(x_2, t_2|x_1, t_1)\{\rho_1(x_1, t_1)/P_1(y_1, t_1)\}\, dx_1$$

Demostración:

$$\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \{\int_{y_1...y_n} \rho_{n+1}(x_{n+1}, t_{n+1;}\, x_n, t_n;...x_1, t_1)\, dx_{n+1}...dx_1\}/P_n(y_n, t_n,...x_1, t_1)$$

$$= \int_{y_1...y_n} \rho_{1|n}(x_{n+1}, t_{n+1}|x_n, t_n;...x_1, t_1)\{\rho_n(x_n,...t_1)/P_n(y_n,...t_1)\}\, dx_n...dx_1$$

$$= \int_{y_1...y_n} \rho_{1|1}(x_{n+1}, t_{n+1}|x_n, t_n)\{\rho_n(x_n,...t_1)/P_n(y_n,...t_1)\}\, dx_n...dx_1$$

Si $n = 1$, resulta la segunda parte del lema. Para $n > 1$ podemos hacer algunos cálculos más:

$$\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \int_{y_1...y_n} \rho_{1|1}(x_{n+1}, t_{n+1}|x_n, t_n)\{\rho_n(x_n,...t_1)/P_n(y_n,...t_1)\}\, dx_n...dx_1$$

$$= \int_{y_n} \rho_{1|1}(x_{n+1}, t_{n+1}|x_n, t_n)\{\int_{y_1...y_{n-1}} \rho_n(x_n,...t_1)dx_{n-1}...dx_1/P_n(y_n,...t_1)\}\, dx_n$$

Pero:

$$\int_{y_1...y_{n-1}} \rho_n(x_n,...t_1)\, dx_{n-1}...dx_1 = \sigma_{1|n-1}(x_n, t_n|y_{n-1},...t_1)P_n(y_{n-1},...t_1)\ y$$

$$P_n(y_n,...t_1) = P_{1|n-1}(y_n, t_n|y_{n-1}...t_1)P_n(y_{n-1},...t_1)$$

de donde resulta la primera parte del lema.

Ahora podemos enunciar condiciones suficientes para asegurar la markovianidad y homogeneidad de la proyección. Aparte de la condición **(P0)**, introduciremos en la siguiente proposición dos condiciones más: una de ellas es una "condición de Markov débil" que involucra a elementos de X y de Y y que conducirá a la condición de Markov en Y, y la segunda es una condición de invariancia en el tiempo que conducirá a la homogeneidad de la proyección.

Proposición 1: Si el proceso estocástico en X cumple las condiciones:

(P1) Para cada $x \in X$, $y' \in Y$ e instantes $t, t' \geq 0$ se verifica:

$\sigma_{1|1}(x, t|y', t')/P_{1|1}(y, t|y', t') = \rho_1(x, t)/P_1(y, t)$, donde $y = \pi(x)$.

(P2) Para cada $x \in X$ y $t \geq 0$ se verifica:

$\rho_1(x, t)/P_1(y, t) = \rho_1(x, 0)/P_1(y, 0)$, donde $y = \pi(x)$.

entonces el proceso proyectado en Y es markoviano y homogéneo.

Demostración:

Probemos por inducción sobre n que $\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \sigma_{1|1}(x_{n+1}, t_{n+1}|y_n, t_n)$, de donde resultará por el lema 1 la markovianidad. Para n = 2, apliquemos el lema 2:

$\sigma_{1|2}(x_3, t_3|y_2, t_2, y_1, t_1) = \int_{y_2} \rho_{1|1}(x_3, t_3|x_2, t_2)\{\sigma_{1|1}(x_2, t_2|y_1, t_1)/P_{1|1}(y_2, t_2|y_1, t_1)\}\, dx_2$

y por **(P1)**, $\sigma_{1|1}(x_2, t_2|y_1, t_1)/P_{1|1}(y_2, t_2|y_1, t_1) = \rho_1(x_2, t_2)/P_1(y_2, t_2)$, ya que $y_2 = \pi(x_2)$.

Por otro lado:

$\sigma_{1|1}(x_3, t_3|y_2, t_2) = \int_{y_2} \rho_{1|1}(x_3, t_3|x_2, t_2)\{\rho_1(x_2, t_2)/P_1(y_2, t_2)\}\, dx_2$

lo cual implica la igualdad para n = 2.

Supongamos que $\sigma_{1|n-1}(x_n, t_n|y_{n-1}, t_{n-1};...y_1, t_1) = \sigma_{1|1}(x_n, t_n|y_{n-1}, t_{n-1})$. Aplicando el lema 1, de esto resulta que $P_{1|n-1}(y_n, t_n|y_{n-1}, t_{n-1};...y_1, t_1) = P_{1|1}(y_n, t_n|y_{n-1}, t_{n-1})$. Por lo tanto:

$\sigma_{1|n-1}(x_n, t_n|y_{n-1},...t_1)/P_{1|n-1}(y_n, t_n|y_{n-1}...t_1) = \sigma_{1|1}(x_n, t_n|y_{n-1}, t_{n-1})/P_{1|1}(y_n, t_n|y_{n-1}, t_{n-1})$

$= \rho_1(x_n, t_n)/P_1(y_n, t_n)$

donde aplicamos nuevamente **(P1)**, recordando que $y_n = \pi(x_n)$. Entonces:

$\sigma_{1|n}(x_{n+1}, t_{n+1}|y_n, t_n;...y_1, t_1) = \int_{y_n} \rho_{1|1}(x_{n+1}, t_{n+1}|x_n, t_n)\{\rho_1(x_n, t_n)/P_1(y_n, t_n)\}\, dx_n$

$= \sigma_{1|1}(x_{n+1}, t_{n+1}|y_n, t_n)$

con lo que se tiene la markovianidad del proceso en Y.

Probemos ahora la homogeneidad de $\sigma_{1|1}$:

$\sigma_{1|1}(x, t|y', t') = \int_{y'} \rho_{1|1}(x, t|x', t')\{\rho_1(x', t')/P_1(y', t')\}\, dx' = \int_{y'} \rho_{1|1}(x, t-t'|x', 0)\{\rho_1(x', 0)/P_1(y', 0)\}\, dx'$

$= \sigma_{1|1}(x, t-t'|y', 0)$

donde se aplicó **(P2)** y que el proceso en X es homogéneo.

La interpretación física de las condiciones **(P1)** y **(P2)** se aprecia claramente cuando son re-escritas para el caso de un espacio de estados discreto En este caso, se tiene que las condiciones **(P1)** y **(P2)** se transforman en:

(P1') $P_{1|2}(x, t|\pi(x), t; y', t') = P_{1|1}(x, t|\pi(x), t)$

(P2') $P_{1|1}(x, t|\pi(x), t) = P_{1|1}(x, 0|\pi(x), 0)$

En efecto, cuando X es discreto, la igualdad $\sigma_{1|1}(x, t|y', t')/P_{1|1}(y, t|y', t') = \rho_1(x, t)/P_1(y, t)$ se transforma en: $P_{1|1}(x, t|y', t')/P_{1|1}(y, t|y', t') = P_1(x, t)/P_1(y, t)$. Pero:

$P_{1|1}(x, t|y', t')/P_{1|1}(y, t|y', t') = \{P_2(x, t; y', t')P_1(y', t')\}/\{P_2(y, t; y', t')P_1(y', t')\} =$

$= P_3(x, t; y, t; y', t')/P_2(y, t; y', t') = P_{1|2}(x, t|y, t; y', t')$

ya que el evento $\{x, t\}$ está contenido en $\{y, t\}$. Usando el mismo argumento:

$P_1(x, t)/P_1(y, t) = P_2(x, t; y, t)/P_1(y, t) = P_{1|2}(x, t|y, t)$

Teniendo en cuenta que $x \in y$, se puede escribir $y = \pi(x)$, y así obtenemos **(P1')**. De la misma similar se comprueba **(P2')**.

(P1') indica que una vez que el sistema está dentro de una clase, la información acerca de las transiciones pasadas es irrelevante. Notemos que esto es más débil que la condición markoviana, en donde en lugar de la condición $\{\pi(x), t\}$ se tiene un evento {y", t"} con y", t" arbitrarios. Si interpretamos las clases de la partición Y como las cuencas atractivas de un potencial definido en X, esta condición nos dice que las transiciones entre cuencas deben ser "uniformes" en el sentido de que no puede haber una transición particular entre cuencas que "impulse" al sistema a una región particular de la nueva cuenca. Por supuesto que esto depende de la escala de tiempo en la que estemos examinando al sistema [2]: si esta es muy pequeña, es posible que seamos capaces de observar las irregularidades de la dinámica microscópica que dan preferencia a determinadas conformaciones cada vez que efectúa un salto entre cuencas.

(P2') nos dice que la distribución de probabilidades dentro de una clase es invariante en el tiempo, una hipótesis que es usualmente utilizada en numerosos modelos matemáticos de sistemas físicos en los cuales la partición corresponde a las cuencas atractivas de un potencial. Por ejemplo, en el llamado modelo de energías aleatorias (REM, Random-Energy Model) [8, 9], la distribución de probabilidades dentro de una cuenca es gaussiana con desvío y media que permanecen constantes a lo largo del tiempo.

2.5 Conmutatividad de procesos estocásticos

En la sección anterior se encontraron condiciones que garantizan que la proyección de un proceso markoviano y homogéneo sea también un proceso markoviano y homogéneo. El problema es que las condiciones enunciadas pueden resultar demasiado restrictivas, en especial **(P0)**, es decir, que todos los caminos en Y tengan probabilidad no nula. Particularmente, la distribución inicial $P_0(y) = P(y, 0)$ debe ser estrictamente positiva. Por otro lado, nuestro objetivo consiste en obtener la probabilidad proyectada desde X resolviendo la ecuación maestra de una cadena de Markov en Y, que no tiene necesariamente que ser la proyección del proceso en X. En otras palabras, estamos interesados en obtener una conmutatividad de procesos estocásticos que podemos expresar como "resolver ecuación maestra en X y luego proyectar la densidad = proyectar el proceso y luego resolver la ecuación maestra discreta".

De hecho, si la condición **(P0)** no se cumpliera, no sería posible siquiera construir las probabilidades de transición de un paso en Y, ya que:

$$P_t(y|y') = P_{1|1}(y, t|y', 0) = \int_{y'} \int_y T_t(x|x')dx \{\rho_0(x')/P_0(y')\} dx' \qquad (7)$$

por lo tanto $P_0(y')$ debe ser no nula para cada clase y'. Esto es una condición que en general no se cumple. Por ejemplo, en el caso del plegamiento, aquellos productos de cuencas que no correspondan a conformaciones random coil tienen probabilidad inicial nula, y por lo tanto no sería posible proyectar el proceso de acuerdo con la definición 3 para obtener una cadena en Y.

Supongamos ahora que las densidades de transición $T_t(x|x')$ cumplen la siguiente condición:

$$\int_y T_t(x|x') dx = \int_y T_t(x|x'') dx \text{ para todo x' y x'' en la misma clase}$$

En este caso, la expresión (7) se reduce a:

$$P_t(y|y') = \int_{y'} \int_y T_t(x|x')\, dx\, \{\rho_0(x')/P_0(y')\}\, dx' = \int_y T_t(x|x')\, dx \int_{y'} \{\rho_0(x')/P_0(y')\}\, dx'$$

$$= \int_y T_t(x|x')\, dx$$

donde x' es un elemento arbitrario de y'. Es decir, desaparece la probabilidad inicial de la expresión de $P_t(y|y')$. Esto conduce a la siguiente definición:

Definición 3: Un proceso markoviano y homogéneo en X con densidades de transición que cumplen:

(C0) $\int_y T_t(x|x')\, dx = \int_y T_t(x|x'')\, dx$ para todo $x', x'' \in X$ tales que $\pi(x') = \pi(x'')$

se dice compatible con la partición Y.

Bajo esta condición, para cada par de estados y, y' en Y se definen las probabilidades de transición $P_t(y|y')$ como:

$$P_t(y|y') = \int_y T_t(x|x')\, dx$$

donde x' es un elemento cualquiera de y'.

Proposición 2: Si el proceso en X es compatible con la partición Y, entonces las probabilidades de transición $P_t(y|y')$ verifican las ecuaciones de Chapman-Kolmogorov.

Demostración: Sea x' un elemento cualquiera de la clase y'. Entonces, aplicando las ecuaciones de Chapman-Kolmogorov en X y la condición **(C0)**, resulta:

$$P_{t+t'}(y|y') = \int_y T_{t+t'}(x|x')\, dx = \int_y \{\int_X T_t(x|x'')T_{t'}(x''|x')dx''\}\, dx = \int_y \{\sum_{y''}\{\int_{y''} T_t(x|x'')T_{t'}(x''|x')dx''\}\}dx =$$

$$= \sum_{y''} \int_y \int_{y''} T_t(x|x'')T_{t'}(x''|x')dx''dx = \sum_{y'} \{\int_y T_t(x|x'')dx\}\{\int_{y''} T_{t'}(x''|x')dx''\} =$$

$$= \sum_{y'} P_t(y|y'')P_{t'}(y''|y')$$

Este resultado nos indica que, a partir de las densidades de transición $T_t(x|x')$ compatibles con Y, podemos definir un proceso markoviano en Y. La probabilidad inicial en Y es simplemente la proyección de la densidad inicial en X:

$$P_0(y) = \int_y \rho_0(x)\, dx \qquad (8)$$

Utilizando la ecuación maestra en X y la densidad inicial $\rho_0(x)$ se obtiene la evolución temporal de la densidad $\rho(x, t)$. La proyección de esta densidad es:

$$P(y, t) = \int_y \rho(x, t)\, dx \qquad (9)$$

Veamos que la probabilidad (9) es la misma que se obtiene con la cadena de Markov en Y definida por las probabilidades de transición $P_t(y|y')$ y la densidad inicial proyectada (8).

Proposición 3: $\int_y \rho(x, t)\, dx = \sum_{y'} P_t(y|y')\{\int_{y'} \rho_0(x')\, dx'\}$

Demostración:

$$\sum_{y'} P_t(y|y')\{\int_{y'} \rho_0(x')\, dx'\} = \sum_{y'} \{\int_y T_t(x|x')dx\}\int_{y'}\rho_0(x')\, dx' = \int_y \{\sum_{y'}\int_{y'} T_t(x|x')\rho_0(x')dx'\}\, dx =$$

$$= \int_y \{\int_X T_t(x|x')\rho_0(x')dx'\}\, dx = \int_y \rho(x, t)\, dx$$

Es decir, si asociamos al proceso markoviano definido en X por $T_t(x|x')$ y $\rho_0(x)$ la cadena de Markov en Y determinada por las probabilidades $P_t(y|y')$ y la distribución $P_0(y)$, la probabilidad que se obtiene al proyectar $\rho(x, t)$ es precisamente la solución de la ecuación maestra en Y.

Podemos resumir todas estos resultados en un teorema que concrete la idea de conmutatividad de procesos estocásticos expuesta al principio de la sección. Para enunciar este teorema, primero hagamos una definición más:

Definición 4: Se definen los siguientes conjuntos:

(a) $M_Y(X)$ el conjunto de procesos markovianos y homogéneos, cada uno definido por la familia de densidades de transición $T_t(x|x')$ y una densidad inicial $\rho_0(x)$, tales que son compatibles con la partición Y.

(b) M(Y) el conjunto de cadenas de Markov homogéneas sobre Y. Cada una de ellas está definida por la familia de probabilidades de transición $P_t(y|y')$ y una distribución inicial $P_0(y)$.

(c) D(X) el conjunto de funciones de densidad $\rho(x, t)$ definidas en X para $t \geq 0$.

(d) D(Y) el conjunto de probabilidades P(y, t) definidas en Y para $t \geq 0$.

Se definen las siguientes aplicaciones:

(d) θ_X: $M_Y(X) \rightarrow D(X)$ y θ_Y: $M(Y) \rightarrow D(Y)$ que asocian a un proceso de Markov la solución de la ecuación maestra respectiva.

(e) La proyección π_D: $D(X) \rightarrow D(Y)$ que a cada densidad en D(X) le asigna una distribución en D(Y) a través de $P(y, t) = \int_y \rho(x, t)\, dx$.

Entonces podemos enunciar el siguiente teorema de conmutatividad:

Teorema 1: Existe una aplicación π_M: $M_Y(X) \rightarrow M(Y)$ tal que el siguiente diagrama conmuta:

$$\begin{array}{ccc} M_Y(X) & \xrightarrow{\theta_X} & D(X) \\ \downarrow{\scriptstyle \pi_M} & & \downarrow{\scriptstyle \pi_D} \\ M(Y) & \xrightarrow{\theta_Y} & D(Y) \end{array}$$

Demostración: La aplicación π_M: $M_Y(X) \rightarrow M(Y)$ es: $\pi_M[T_t(x|x'), \rho_0(x)] = [P_t(y|y'), P_0(y)]$, con $P_t(y|y') = \int_y T_t(x|x')dx$ y $P_0(y) = \int_y \rho_0(x)dx$. La condición de consistencia **(C0)** y la proposición 2 muestran que π_M está bien definida, mientras que la igualdad $\pi_D \circ \theta_X = \theta_Y \circ \pi_M$ está dada por la proposición 3.

Hagamos un par de observaciones importantes. La primera de ellas es que el proceso en X puede ser consistente con la partición, pero esto no asegura que su proyección sea markoviana ni homogénea. La condición **(C0)** no implica, en particular, **(P0)**. Es decir, el proceso en Y que se asocia al proceso original en X a través de la aplicación π_M no es en general la proyección del proceso original.

En segundo lugar, como Y es numerable, cada cadena de Markov definida en Y tiene asociada una matriz de velocidades de transición que denotamos $W = (W_{yy'})$. Se tienen las siguientes identidades [3]:

$$-W_{yy'}/W_{y'y'} = P(\text{adoptar la conformación } y|\text{se abandona la conformación } y'),\ y \neq y'$$

$$-1/W_{y'y'} = E(\text{tiempo de permanencia en la conformación } y') \qquad (10)$$

La cadena definida por la matriz de velocidades W puede visualizarse de esta manera: el sistema permanece en cada estado y' un tiempo aleatorio que sigue una distribución exponencial de parámetro $-1/W_{y'y'}$, y una vez que abandona el estado y' selecciona un nuevo estado distinto de y' de acuerdo con las probabilidades $-W_{yy'}/W_{y'y}'$.

2.6 La validez de la conmutatividad en el contexto del plegamiento

El teorema de la conmutatividad de procesos estocásticos depende de la validez de la condición **(C0)**. Desde un punto de vista físico, es importante saber que sucede cuando esta condición se cumple de una manera aproximada.

Una situación particularmente importante es la siguiente: $\int_y T_t(x|x')\,dx$ es independiente de x' a partir de cierto instante $t_{y'}$. Es este caso, si definimos:

$$P_t(y|y') = \int_y T_t(x|x')\,dx \text{ si } t \geq t_{y'}$$

$$= P_{ty'}(y|y') \text{ si } t < t_{y'}$$

podemos escribir, sin pérdida de generalidad:

$$\int_y T_t(x|x')\,dx = P_t(y|y') + o(y', x', t)$$

donde o(y', x', t) es una cierta función que cumple o(y', x', t) = 0 para $t \geq t_{y'}$.

Calculemos la evolución de la probabilidad $P(y, t) = \int_y \rho(x, t)\,dx$:

$$P(y, t) = \int_y \rho(x, t)\,dx = \int_y \{\int_X T_t(x|x')\rho_0(x')\,dx'\}\,dx = \sum_{y'} \{\int_{y'}\int_y T_t(x|x')\rho_0(x')\,dxdx'\} =$$

$$= \sum_{y'} \{\int_{y'}\{P_t(y|y') + o(y', x', t)\}\,\rho_0(x')\,dx'\} =$$

$$= \sum_{y'} P_t(y|y')P_0(y') + \sum_{y'} \{\int_{y'} o(y', x', t)\rho_0(x')\,dx'\}$$

Si Y es un conjunto finito, podemos considerar $t_{max} = \max_{y'}\{t_{y'}\}$. Entonces, si $t \geq t_{max}$ resulta que o(y', x', t) = 0 para todo y', luego:

$$P(y, t) = \sum_{y'} P_t(y|y')P_0(y')$$

es decir, la evolución de la probabilidad a partir de t_{max} es la que corresponde cuando se verifica la condición **(C0)**.

Si todos los $t_{y'}$ son muy pequeños en comparación con los escalas de tiempo que deseamos observar, podemos aceptar a efectos prácticos que la condición **(C0)** es válida. Físicamente, los tiempos $t_{y'}$ son los tiempos de equilibración o termalización dentro [7, 8] de las regiones y'.

En el caso del plegamiento, sabemos que una proteína se termaliza [8, 9] dentro de las cuencas de Ramachandran en escalas de tiempo mucho más cortas que los tiempos en los cuales ocurren las transiciones entre cuencas. En consecuencia, podemos aceptar la validez del diagrama conmutativo del teorema 1 para el problema del plegamiento. Esto nos asegura que podemos encontrar una cadena de Markov en el espacio de producto de cuencas Y tal que genere la misma dinámica que obtendríamos al proyectar la solución de la ecuación maestra (2). Esta conclusión es precisamente la que convalida el modelo simplificado que hemos debido introducir para asegurar que el problema del plegamiento tenga un tratamiento computacional accesible.

2.7 El plegamiento como un proceso markoviano discreto: el planteo genérico

Hemos visto que es posible representar la dinámica de plegamiento coma cadena de Markov en el espacio finito formado por los productos de cuencas . Como sabemos, una cadena de Markov queda unívocamente determinada por su matriz de velocidades instantáneas de transición $W = (W_{yy'})$. El objetivo de esta sección es encontrar la matriz de velocidades de transición para el plegamiento.

Consideremos una transición $y' = (y'_1, y'_2,...y'_N) \rightarrow y = (y_1, y_2,...y_N)$ en Y. Esta transición corresponde a las transiciones locales $y'_i \rightarrow y_i$ para aquellas unidades tales que $y'_i \neq y_i$, mientras

que las demás unidades permanecen en sus cuencas. Definamos los conjuntos I y J formados por unidades de la cadena como:

$I = \{i: y'_i \neq y_i\}$, $J = \{i: y'_i = y_i\}$

y definamos los siguientes eventos:

A_i = {la unidad i adopta la cuenca de Ramachadran y_i}

B_i = {la unidad i abandona la cuenca de Ramachadran y'_i}

A = {la proteína adopta la conformación y}

B = {la proteína abandona la conformación y'}

Si $W = (W_{yy'})$ representa la matriz de velocidades de transición en Y, se tiene que, de acuerdo con (10):

$$-W_{yy'}/W_{y'y'} = P(A|B) = P(A, B)/P(B)$$

A continuación obtendremos expresiones para P(A, B) y P(B), para lo cual utilizaremos el hecho de que las dinámicas locales en cada cuenca de Ramachandran son independientes entre sí durante la transición $y' \rightarrow y$. De esta manera, P(A, B) adopta la siguiente forma multiplicativa:

$$P(A, B) = P(\cap_{i\in I}\{\text{ocurren } A_i \text{ y } B_i\})P(\cap_{i\in J}\{\text{no ocurre } B_i\})$$

$$= \{\prod_{i\in I} P(A_i, B_i)\}\{\prod_{i\in J}\{1 - P(B_i)\}\}$$

$$= \{\prod_{i\in I} P(A_i|B_i)P(B_i)\}\{\prod_{i\in J}\{1 - P(B_i)\}\}$$

Por otro lado, el evento complementario a B consiste en que ninguna unidad en y' cambie de cuenca. Aplicando nuevamente la independencia de las dinámicas locales, se tiene que:

$$P(B) = 1 - \prod_{i=1...N}\{1 - P(B_i)\}$$

Por otro lado, $-1/W_{y'y'}$ es el tiempo de permanencia esperado en el estado y'. Podemos aproximar al proceso de abandonar al estado y' como una sucesión de intentos en cada uno de los

cuales la probabilidad p de abandonar y' es p = P(B). El número esperado de intentos hasta que finalmente se abandona el estado y' es 1/p, por lo tanto, si τ es el tiempo que transcurre entre dos intentos consecutivos, el tiempo esperado es τ/P(B), y en consecuencia:

$$-1/W_{y'y'} \approx \tau/P(B)$$

El tiempo característico τ que aparece en esta expresión se puede calcular como el período medio de salto entre cuencas de Ramachandran. Realizaremos este cálculo en el capítulo 4.

Utilizando las expresiones obtenidas hasta ahora, llegamos a que los elementos de la matriz W son:

$$W_{yy'} = \tau^{-1}\{\prod_{i\in I} P(A_i|B_i)P(B_i)\}\{\prod_{i\in J}\{1 - P(B_i)\}\}$$

$$W_{y'y'} = \tau^{-1}\{1 - \prod_{i=1\ldots N}\{1 - P(B_i)\}\} \qquad (11)$$

De esta manera vemos que la cadena queda totalmente caracterizada por las probabilidades $P(A_i|B_i)$ y $P(B_i)$ y el tiempo característico de transición τ.

La probabilidad de adoptar la cuenca y_i dado que la cuenca y'_i es abandonada está dada por el área lacunar (entropía microcanónica) de la cuenca y_i. Es decir:

$$P(A_i|B_i) = A(y_i)/\sum_{y''_i \neq y'_i} A(y''_i) \qquad (12)$$

donde $A(y_i)$ = área lacunar de la cuenca de Ramachandran y_i.

Denotemos con $\Delta G(i, y')$ la disminución en un potencial termodinámico (energía libre) que se asocia con el cambio de cuenca en la unidad i-ésima cuando el sistema completo ocupa inicialmente el estado y'. El cálculo de $\Delta G(i, y')$ requiere, por una parte, que encontremos una conformación torsional $x^* \in y'$ que optimice el potencial efectivo U^{eff}. De esta manera incorporamos el hecho de que la proteína explora muy rápidamente los mínimos del potencial U^{eff} que son compatibles con la asignación de cuencas definida por la LTM y'. En consecuencia,

la pérdida $\Delta G(i, y')$ será calculada con respecto a la conformación óptima x*. Por otra parte, para obtener $\Delta G(i, y')$ será necesario también evaluar la pérdida entrópica conformacional asociada con mantener al residuo i en la LTM y'. Esto será descripto en detalle en el capítulo 4.

Una vez que tenemos la cantidad $\Delta G(i, y')$, podemos calcular $P(B_i)$ como sigue:

$$P(B_i) = \exp[\kappa \Delta G(i, y')] \quad \text{si} \quad \Delta G(i, y') \leq 0$$

$$P(B_i) = 1 \qquad \text{si} \quad \Delta G(i, y') > 0 \qquad (13)$$

donde $\kappa = 1/RT$, siendo T la temperatura absoluta y R la constante de los gases ideales.

El cómputo de estas probabilidades guarda un directo correlato con el dictado de la intuición: si un caminante tiene que caer en uno de entre varios pozos posibles de un terreno plano, la probabilidad de caer en uno dado depende de su área lacunar y no de su profundidad, mientras que si se encuentra en un pozo la probabilidad de salir de él depende de la profundidad del mismo.

2.8 Representación algorítmica del plegamiento basada en el planteo genérico anterior

Si bien la cadena de Markov definida en Y en la sección anterior representa rigurosamente la dinámica discreta del sistema, su utilización es impráctica debido a que la matriz $W = (W_{yy'})$ tiene aproximadamente 3^{2N} elementos. Por lo tanto, lo que hacemos en realidad es generar trayectorias individuales en Y que correspondan a realizaciones particulares del proceso determinado por la ecuación maestra $dP/dt = WP$. Denotemos con $\varphi_k : [0, t_{total}] \rightarrow Y$ a una trayectoria individual. Entonces podemos aproximar a $P_y(t)$ como

$$P_y(t) \approx \#\{k: \varphi_k(t) = y\}/M$$

donde M es el número total de trayectorias generadas, $1 \leq k \leq M$. La construcción de $P_y(t)$ a través de este método es posible en nuestro modelo ya que los caminos discretos de plegamiento son fácilmente computables.

Ahora expondremos el algoritmo de plegamiento que nos permite generar trayectorias individuales discretas. El algoritmo consiste en generar de manera iterativa para cada unidad i transiciones entre las cuencas de Ramachandran de acuerdo con las probabilidades $P(B_i)$ y $P(A_i|B_i)$: primero se calcula $P(B_i)$ en función de la conformación actual del sistema, luego aquellas unidades que son elegidas para cambiar de cuenca de acuerdo con $P(B_i)$ seleccionan una nueva cuenca utilizando $P(A_i|B_i)$. El intervalo de tiempo de estas iteraciones corresponde al tiempo característico τ. El pseudocódigo del algoritmo para generar una trayectoria $\varphi(t)$ en el intervalo $[0, t_{total}]$ es el siguiente:

```
Comienzo
  t:= 0
  Seleccionar φ(0) en el ensamble random coil
  Mientras t ≤ t_total repetir
  Comienzo
    De i:= 1 a N hacer
    Comienzo
      Calcular ΔG(i,φ(t))
      p_i:= exp[ΔG(i,φ(t))/RT]
    Fin
    De i:= 1 a N hacer
    Comienzo
      r:= número aleatorio uniforme en [0, 1]
      Si r < p_i entonces
        φ(t + τ)_i:= nueva cuenca de Ramachandran seleccionada
                     con las probabilidades A(y_i)/∑_{y'_i≠φ(t)_i} A(y'_i)
      Sino
        φ(t + τ)_i:= φ(t)_i
    Fin
    t:= t + τ
  Fin
Fin
```

2.9 Referencias

[1] *Foundations of Mechanics*. R. Abraham, J. E. Marsden. The Benjamin/Cummings Publishing Company (1978).

[2] *Stochastic Processes in Physics and Chemistry.* N. G. Van Kampen. North-Holland (1981).

[3] *Probability and Random Processes.* G. R. Grimmet, D. R. Stirzaker. Clarendon Press (1991).

[4] *Markov Chains with Stationary Transition Probabilities*. N. L. Chung. North-Holland (1967).

[5] A geometric framework for polymer folding. H. Cendra, A. Fernández, W. Reartes. *J. Math. Chem.* **19**, 331-336 (1996).

[6] Macromolecular conformational dynamics in torsional angle space. S. He, H. A. Scheraga. *J. Chem. Phys.* **108**, 271-286 (1997).

[7] Realistic master equation modeling of relaxation on complete potential energy surfaces: kinetic results. K. D. Ball, R. S. Berry. *J. Chem. Phys.* **109**, 8557-8571 (1998).

[8] Random-energy model: An exactly solvable model of disordered systems. B. Derrida. *Phys. Rev. B* **24**, 2613-2618 (1981).

[9] How large should proteins be? The minimal size of a good structure seeker. A. Fernández, A. Colubri, T. Burastero, A. Tablar. *Phys. Chem. Chem. Phys.* **1**, 4347-4354 (1999).

Capítulo 3

Diseño de un potencial efectivo en T^{2N}

3.1 Introducción y motivaciones: la proteína como generador de su propio ambiente

Como se indicó en el capítulo anterior, una vez que el sistema ocupa una combinación de cuencas particular o LTM, es necesario encontrar una conformación torsional (ϕ, ψ) que, sujeta a las restricciones impuestas por la LTM, optimice el potencial efectivo U^{eff} definido en $X = T^{2N}$. Esta conformación es obtenida a través de una optimización Monte-Carlo que se describirá en el capítulo siguiente. De esta manera, es posible asociar un patrón óptimo de interacciones de largo rango a una LTM determinada.

Este potencial efectivo U^{eff} debe cumplir dos objetivos: **(a)** representar con realismo las diferentes interacciones intramoleculares y el balance energético entre ellas, y **(b)** incorporar implícitamente el efecto del solvente en las interacciones entre los residuos de la proteína. Para alcanzar (**a)**, hemos parametrizado los diferentes términos de nuestra función potencial (Tablas IIa-c), de manera tal que hay suficiente latitud para ajustar los parámetros y encontrar una combinación satisfactoria. Con relación a (**b)**, el efecto del solvente fue introducido operacionalmente en el potencial intramolecular a través de una atracción hidrofóbica efectiva, una repulsión hidrofóbico-polar de dos cuerpos, y un re-escalamiento de todas las contribuciones de a pares de acuerdo con los niveles de desolvatación de los residuos interactuantes, tal como es explicado en la siguiente sección.

La imagen que resulta de nuestro modelo revela que, a medida que avanza el plegamiento, la energía hidrofóbica es convertida progresivamente en energía efectiva de

desolvatación del backbone, volviendo en determinantes de la estructura de la proteína a los puentes de hidrógeno amida-carbonilo intramoleculares (estos son los que se forman entre los grupos NH y CO de aminoácidos diferentes). Esta consecuencia es muy importante ya que permite solucionar una contradicción que subsiste desde hace tiempo: por un lado, los puentes de hidrógeno amida-carbonilo intramoleculares son fundamentales para determinar la geometría torsional de las estructuras secundarias básicas (hojas, hélices y turns), mientras que por el otro lado, la proteína puede formar puentes de hidrógeno mucho más fácilmente con el agua circundante. En particular, el escenario que surge para plegadores de dos estados es el siguiente: el plegamiento está dirigido por una búsqueda de una topología nucleante que garantice la supervivencia de los puentes de hidrógeno intramoleculares del ataque del agua.

Asumiremos que U^{eff} puede expresarse como la suma siguiente $U^{eff} = \sum_{i,j} U^{eff}(i, j)$, donde $U^{eff}(i, j)$ representa la energía efectiva de interacción entre los residuos i y j. Sin embargo, esto no quiere decir que $U^{eff}(i, j)$ dependa solamente de la distancia entre i y j. Precisamente, la incorporación implícita del solvente tiene como consecuencia que las interacciones de dos cuerpos $U^{eff}(i, j)$ estén re-escaladas por la proximidad de otros residuos de la cadena. De este aspecto nos ocuparemos en la sección siguiente.

A efectos de fijar la terminología que utilizaremos de aquí en adelante, diremos que un par (i, j) está formando un contacto si su energía efectiva de interacción es negativa: $U^{eff}(i, j) < 0$. Por el contrario, (i, j) está formando un anticontacto cuando $U^{eff}(i, j) > 0$. Un contacto (o anticontacto) (i, j) se dice de largo rango si $|j-i| > 10$.

3.2 Re-escalamiento de las contribuciones de dos cuerpos dependiente de la desolvatación

Modelamos las interacciones de largo rango entre los residuos introduciendo los siguientes términos en nuestro potencial efectivo [1]:

$$U^{eff} = U_{EV} + U_{solv} + U_{ion} + U_{dip} + U_{Hbond} + U_{SS}$$

donde: U_{EV} representa una contribución de volumen excluido que tiene en cuenta la repulsión estérica entre las cadenas laterales, U_{solv} es el término solvofóbico efectivo que representa la atracción entre residuos hidrofóbicos y la repulsión entre residuos hidrofóbicos y polares, U_{ion} es la energía de interacción iónica entre las cadenas laterales cargadas, U_{dip} modela las interacciones dipolo-dipolo del backbone, U_{Hbond} corresponde a los puentes de hidrógeno del backbone y, finalmente, U_{SS} mide la energía de los puentes disulfuro que pueden aparecer entre un par de cisteínas. Esta función está basada en un potencial semi-empírico que ha sido testeada exitosamente en trabajos anteriores [1, 2, 3].

En una aproximación de orden cero, denotada genéricamente U^0, cada uno de estos términos puede ser expresado como una suma sobre contribuciones de a pares: $U^0 = \sum_{i,j} U^0(i, j)$. En esta aproximación, la función potencial no refleja el efecto que los ambientes locales de solvente tienen sobre la estabilidad de las interacciones dieléctrico-dependientes (solvofóbica, iónica y puente de hidrógeno). Para incorporar este efecto, re-escalamos la contribución de orden cero de cada par (i, j), $U^0(i, j)$, introduciendo factores de renormalización f_i y f_j, los cuales dependen del nivel de desolvatación de los residuos i y j. De esta manera, la energía re-escalada del par (i, j) es $U(i, j) = f_i f_j U^0(i, j)$, donde $f_i = f_i (L_i)$ y L_i = nivel de desolvatación del residuo i.

Para cada residuo i, definimos la variable L_i como $L_i = V_H(i)/V_T$, donde V_T es el volumen de una esfera S_i de radio igual a 7 Å, centrada en el carbono α i-ésimo, y $V_H(i)$ es el volumen de

esta esfera ocupado por cadenas laterales hidrofóbicas que se encuentran suficientemente cerca de i para ser incluidas en la esfera S_i. De esta definición, se sigue que L_i varía en el intervalo [0, 1] y claramente refleja el nivel local de enterramiento hidrofóbico del residuo. En este modelo, $L_i \approx 0$ significa que el residuo i está completamente expuesto al solvente, mientras que $L_i \approx 1$ significa que el residuo i está totalmente protegido.

La dependencia funcional de los factores f_i en los niveles de desolvatación L_i modela cuantitativamente la manera en que el solvente afecta la estabilidad efectiva de las interacciones dieléctrico-dependientes. Por ejemplo, el efecto de desolvatar un puente de hidrógeno es que el mismo se vuelve efectivamente más fuerte (a pesar de que su constante de fuerza es afectada exclusivamente por la orientación relativa de los residuos involucrados) con respecto a un puente expuesto al solvente. Mientras tanto, la atracción hidrofóbica que aparece en un par desolvatado se vuelve efectivamente más débil de acuerdo con el entierro relativo de la superficie expuesta al solvente: el efecto hidrofóbico entrópico es inexistente en la ausencia de solvente en las regiones circundantes.

Las contribuciones de orden cero, los parámetros ajustables y los factores de renormalización f_i (cuando corresponda) para cada término de la función potencial serán discutidos en las siguientes secciones.

3.3 Potencial de volumen excluido

Este término representa la repulsión estérica entre las cadenas laterales. Ya que en nuestro modelo las cadenas laterales son elipsoides rígidos e impenetrables, la energía de volúmen excluido entre las unidades i y j es infinita cuando los elipsoides se intersectan, mientras que es 0 en caso contrario.

En el caso particular de que las cadenas laterales i y j son esferas con radios R_i y R_j, respectivamente, centradas cada una de ellas en los carbonos α correspondientes, entonces podemos escribir:

$$U_{EV}(i, j) = \infty \text{ si } r_{ij} \leq R_i + R_j$$

$$U_{EV}(i, j) = 0 \text{ si } r_{ij} > R_i + R_j$$

donde r_{ij} es la distancia entre los carbonos α i y j.

La implementación de nuestro algoritmo de plegamiento permite incorporar dependencias más elaboradas del potencial U_{EV} en función de la distancia r_{ij}. Es posible, por ejemplo, modelar un potencial Lennard-Jones del tipo 6-12. Sin embargo, el nivel de resolución estructural con el cual estamos representando a las cadenas laterales es incompatible con un potencial 6-12: éste es adecuado para cuantificar la repulsión estérica entre átomos, pero notemos que aquí estamos modelando a las cadenas laterales simplemente como esferas que representan el volumen promedio ocupado por sus átomos.

3.4 Potencial solvofóbico efectivo

El potencial solvofóbico efectivo puede ser expresado como la suma de un término atractivo y un término repulsivo: $U_{solv} = U_{solv,\,hh} + U_{solv,\,hp}$. La atracción solvofóbica h-h se debe a la necesidad de minimizar la pérdida entrópica asociada con el ordenamiento del solvente alrededor de las cadenas laterales hidrofóbicas, mientras que la interacción h-p entre una cadena lateral hidrofóbica y una cadena lateral polar es energéticamente desfavorable porque las capas de solvente generadas alrededor de cada residuo son incompatibles entre sí debido a la diferente orientación relativa de los dipolos del solvente.

Las contribuciones de pares de ambos términos, atractivo y repulsivo, son modeladas de la siguiente manera:

$$U^0_{solv,\ hh}(i, j) = c_{hh}(i)\ c_{hh}(j)\ F_{hh}(r_{ij})$$

$$U^0_{solv,\ hp}(i, j) = c_{hp}(i)\ c_{hp}(j)\ F_{hp}(r_{ij})$$

donde r_{ij} es la distancia entre los carbonos α i y j. En el caso de la atracción solvofóbica, $F_{hh}(r_{ij})$ representa una foso energético con una profundidad de -3,1 kcal/mol, que comienza en 4,5 Å y finaliza en 6,7 Å [1]. El valor de la constante $c_{hh}(i)$ es proporcional al área superficial efectiva de la cadena lateral del residuo i. Los valores adoptados son mostrados en la Tabla IIa. Para la repulsión solvofóbica, la función $F_{hp}(r_{ij})$ representa una elevación en el potencial con una altura de 1 kcal/mol, que comienza en 3,5 Å y finaliza en 6,7 Å. En el caso de la repulsión h-p, si asumimos que i es hidrofóbico y que j es polar, entonces $c_{hp}(i) = c_{hh}(i)$, mientras que $c_{hp}(j)$ es proporcional a la polaridad del residuo j. Los valores adoptados para estas últimas constantes son mostrados en la Tabla IIb. Los rangos de ambos términos fueron ajustados observando las distancias entre los carbonos α en las estructuras secundarias típicas: para las hélices α, estas distancias varían entre 6,2 y 6,5 Å, mientras que en las hojas β comienzan en 4,6 Å y finalizan en 6 Å (aproximadamente). Por otro lado, hay que notar que una inspección de las interacciones terciarias de proteínas naturales arbitrarias indica que las distancias entre residuos hidrofóbicos o bien entre residuos hidrofóbicos y polares pertenecen a los rangos recién señalados. Vemos que, tal como ha sido construido, el potencial solvofóbico captura los rangos de interacción de hélices α, hojas β e interacciones terciaras genéricas. El término solvofóbico repulsivo comienza en distancias menores para evitar la proximidad de aminoácidos con distinta preferencia por el solvente.

El factor de renormalización del potencial solvofóbico fue construido como una función que decrece monotónicamente de 1 a 0 a medida que el nivel de desolvatación varía de 0,7 a su valor máximo 1, reflejando de esta manera el hecho de que la atracción y la repulsión solvofóbica se debilitan progresivamente a medida que la cantidad de solvente que rodea a los residuos disminuye. Estos parámetros son consistentes con los tiempos de supervivencia típicos de estructuras con diferentes niveles de exposición al solvente [1, 2, 4-11].

	a	b	c	d
i	$c_{hh}(i)$	$c_{hp}(i)$	q(i)	z(i)
ALA	0.5	0.5	0.0	0
ARG	1.1	1.1	+1.0	5
ASN	0.9	0.9	+0.4	2
ASP	0.9	1.1	-0.8	2
CYS	1.0	1.0	0.0	1
GLN	1.0	0.9	+0.4	3
GLU	1.0	1.0	-0.8	3
GLY	0.5	0.5	0.0	0
HIS	1.0	1.0	+0.8	2
ILE	1.1	1.1	0.0	2
LEU	1.1	1.1	0.0	2
LYS	1.0	1.1	+1.0	4
MET	1.0	1.1	0.0	3
PHE	1.2	1.2	0.0	2
PRO	0.8	0.8	0.0	0
SER	0.6	0.6	0.0	1
THR	0.4	0.4	0.0	1
TRP	1.4	1.4	0.0	2
TYR	1.0	1.0	0.0	2
VAL	1.1	1.1	0.0	1

Tabla I:
Tabla IIa: Constantes de escala multiplicativas c_{hh} para cada aminoácido asociadas con el potencial solvofóbico atractivo.
Tabla IIb: Constantes de escala multiplicativas c_{hp} para cada aminoácido asociadas con la contribución solvofóbica repulsiva.
Tabla IIc: Cargas efectivas de orden cero para cada aminoácido a pH 7.
Tabla IId: Número de diedros libres presentes en las cadenas laterales de cada aminoácido.

3.5 Potencial iónico efectivo dependiente del contexto

La interacción iónica efectiva entre dos residuos cargados en el vacío es representada por una función $U^0_{ion}(i, j) = q(i)\, q(j)\, F_{ion}(r_{ij})$, donde r_{ij} es la distancia entre los carbonos α i y j, q(i) y q(j) son las cargas efectivas adimensionales de los aminoácidos y:

$$F_{ion}(r_{ij}) = c_{ion}/ r_{ij}$$

c_{ion} es un parámetro ajustable cuyo valor fue fijado en 7,5 kcal/mol Å. La Tabla IIc muestra la carga efectiva de los aminoácidos en pH neutral, sin ninguna renormalización debida a la conformación local de la cadena. Estas conformaciones locales pueden alterar dramáticamente el pKa efectivo de los donantes o aceptores de protones y, por consiguiente, las cargas efectivas. Es por esto que los valores mostrados en la tabla deben ser interpretados como promedios de ensamble.

Este potencial es re-escalado de acuerdo con los niveles de desolvatación de los residuos, ya que la interacción iónica claramente depende en dieléctrico efectivo del medio circundante. Las moléculas de solvente tienden a ocultar las cargas de las cadenas laterales, por este motivo utilizamos un factor de renormalización que aumenta desde 0,5 a 1,2 a medida que el nivel de desolvatación varía de 0 a 0,8. Los valores adoptados están en concordancia con datos experimentales reunidos acerca de la estabilidad de puentes iónicos en dependencia con el contexto y la protección de interacciones iónicas [1, 4-9, 12, 13].

3.6 Potencial dipolo-dipolo

Este potencial representa la interacción entre los dipolos de los grupos amida de cada aminoácido y ha sido modelada utilizando la expresión rigurosa para la energía de interacción entre dos dipolos [14].

Sean $\mathbf{m}_i$ y $\mathbf{m}_j$ los vectores dipolares normalizados de los residuos i y j, y $\mathbf{m}_{ij}$ el vector que une a estos dipolos. Entonces la energía dipolar es:

$$U^0_{dip}(i, j) = c_{dip} \{(\mathbf{m}_i \cdot \mathbf{m}_j) / \mu_{ij}^3 - 3 (\mathbf{m}_i \cdot \mathbf{m}_{ij})(\mathbf{m}_j \cdot \mathbf{m}_{ij}) / \mu_{ij}^5 \}$$

donde el parámetro c_{dip} se ajustó en 7 kcal/mol Å ^{3}.

A pesar de que la interacción dipolar es dependiente del dieléctrico, no la hemos re-escalado en función de su medio circundante. Esto se debe a su dependencia cúbica en la distancia de separación entre los dipolos: esta contribución es esencialmente de corto rango y por lo tanto no puede ocasionar un reacomodamiento significativo del solvente. De hecho, el principal efecto de la interacción dipolo-dipolo consiste en el relajamiento local de la cadena hacia la conformación extendida de β-strand, la cual permite el alineamiento favorable (anti-paralelo) de dipolos consecutivos, un efecto local que no está mediatizado por el solvente.

3.7 Potencial puente de hidrógeno efectivo

El energía de puente de hidrógeno de orden cero entre los residuos i y j depende de la orientación relativa de los residuos y es modelada a través de la siguiente expresión:

$$U^0_{Hbond}(i, j) = E_{Hbond}(H_i, O_j) + E_{Hbond}(H_j, O_i)$$

que tiene en cuenta los dos puentes de hidrógeno del backbone que pueden ser formados *a-priori* entre los residuos i y j. La función $E_{Hbond}(H_i, O_j)$ toma la forma del siguiente producto:

$$E_{Hbond}(H_i, O_j) = G_{Hbond}(H_i, O_j) F_{Hbond}(H_i, O_j)$$

donde el factor $G_{Hbond}(H_i, O_j)$ penaliza la no linealidad del puente de hidrógeno, mientras que $F_{Hbond}(H_i, O_j)$ mantiene la distancia H_i-O_j entre 1 Å y 3 Å. En el caso de un puente de hidrógeno geométricamente perfecto, tenemos $E_{Hbond}(H_i, O_j)$ = -1,1 kcal/mol.

Utilizamos la fórmula de Pauling para $G_{Hbond}(H_i, O_j)$ [15]. Sea δ_{ij} el ángulo formado entre los vectores N_i—H_i y H_i—O_j. Entonces la fórmula es:

$$G_{Hbond}(H_i, O_j) = h + c\ \delta_{ij}^2 \text{ si } \delta_{ij}^2 \leq -h/c$$

$$G_{Hbond}(H_i, O_j) = 0 \qquad \text{si } \delta_{ij}^2 > 0$$

en donde h = -1,1 kcal/mol y la constante $c > 0$ modula la tolerancia del puente de hidrógeno a las distorsiones geométricas.

La habilidad del agua para desplazar a los puentes de hidrógeno intramoleculares está bien establecida [1]. En consecuencia, la supervivencia de un puente de hidrógeno intramolecular requiere que la cadena encuentre una topología protectiva que inhiba la distorsión de la estructura del solvente circundante, y por lo tanto, previniendo el ataque del agua [1, 5]. Basados en este hecho, hemos introducido un factor de renormalización de la energía efectiva de un puente de hidrógeno que aumenta linealmente de 0 a 2 a medida que el nivel de desolvatación va de 0 a 0,7, permaneciendo constante hasta que el nivel de desolvatación alcanza el máximo. La justificación de esta definición operacional está basada en datos de intercambio de protones que implican que la energía efectiva de un puente de hidrógeno intramolecular expuesto al solvente es nula. Esto se debe a que la constante aparente de equilibrio entre la conformación con puentes intramoleculares y la conformación con puentes formados con el agua es muy próxima a cero cuando es normalizada con respecto al equilibrio asociado con exponer (u ocultar) al backbone del solvente [1, 5, 11, 12].

3.8 Potencial disulfuro

Este potencial modela la única interacción covalente de largo rango en una proteína, que es la unión entre dos cisteínas. La naturaleza covalente de esta interacción hace que sea extremadamente fuerte, lo cual condiciona de manera decisiva la dinámica de plegamiento en proteínas que poseen más de una cisteína. En estas circunstancias, el plegamiento resulta ser un proceso en donde las transiciones están dictadas por las recombinaciones en el patrón de puentes disulfuros [14], las cuales ocurren mucho más lentamente que los cambios en las restantes interacciones.

Nuestro potencial disulfuro tiene los mismos rangos que el potencial solvofóbico atractivo descripto anteriormente, si bien el pozo de potencial en este caso es más profundo, a fin de representar la mayor fuerza de esta interacción.

3.9 Referencias

[1] Conformational-dependent environments in folding proteins. A. Fernández. *J. Chem. Phys.* **114**, 2489-2502 (2001).

[2] Topology to geometry in protein folding: b-lactoglobulin. A. Fernández, A. Colubri, R. S. Berry. *Proc. Natl. Acad. Sci. USA* **97**, 14062-14066 (2000).

[3] Finding the collapse-inducing nucleus in a folding protein. A. Fernández, G. Appignanesi, A. Colubri. *J. Chem. Phys* **114**, 8678-8684 (2001).

[4] Self-organization and mismatch tolerance in protein folding. General theory and an application. A. Fernández, R. S. Berry. *J. Chem. Phys.* **112**, 5212-5222 (2000).

[5] Cooperative walks in a cubic lattice: Protein folding as a many-body problem. A. Fernández. *J. Chem. Phys.* **115**, 7293-7297 (2001).

[6] Is protein folding hierarchic? I. R. Baldwin, G. D. Rose. *Trends Biochem. Sci.* **24**, 26-76 (1999).

[7] Effective dynamics in the space of contact maps. M. Vendruscolo, E. Domany. *Fold. Des.* **3**, 329-336 (1998).

[8] Pairwise contact potentials are unsuitable for protein folding. M. Vendruscolo, E. Domany. *J. Chem. Phys.* **109**, 11101-11108 (1998).

[9] Context-dependent secondary structure formation of a designed protein sequence. D. L. Minor, P. S. Kim. *Nature* **380**, 730-734 (1996).

[10] Microscopic dynamics from a coarsely defined solution of the protein folding problem. A. Fernández, A. Colubri. *J. Math. Phys.* **39**, 3167-3187 (1998).

[11] Nucleation theory for helix dismantling in protein folding. A. Fernández, A. Colubri. *Phys. Rev. E* **60**, 4645-4651 (1999).

[12] Contribution of a buried hydrogen bond to lambda repressor folding kinetics. J. K. Myers, T. G. Oas. *Biochemistry* **38**, 6761-6768 (1999).

[13] Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. S. Takada, Z. Luthey-Schulten, P. G. Wolynes. *J. Chem. Phys.* **110**, 11616-11629 (1999).

[14] *Biophysical Chemistry, Part I: The Conformation of Biological Macromolecules.* C. R. Cantor, P. R. Schimmel. W. H. Freeman and Company (1980).

[15] *The Nature of the Chemical Bond*. L. Pauling. W. H. Freeman and Company (1954).

Capítulo 4

Diseño de un algoritmo de plegamiento capaz de abarcar tiempos biológicamente relevantes

4.1 Descripción general del algoritmo: la necesidad de comprometer el detalle estructural

Como ya hemos indicado previamente, nuestro enfoque *ab-initio* está basado en la generación de caminos de plegamiento definidos sobre un granulado grueso del espacio de conformaciones, lo cual vuelve computacionalmente accesibles las escalas de tiempo características del plegamiento y permite además hacer predicciones estructurales. En contraste con la mayoría de las herramientas utilizadas para estudiar trayectorias de plegamiento [1-3], nuestro algoritmo no utiliza información *a-priori* del plegamiento final, o potenciales de tipo Gô que penalizan las conformaciones que se apartan de la estructura nativa. El algoritmo ocasionalmente reduce la resolución estructural para hacer accesibles las escalas de tiempo biológicamente relevantes. A efectos de determinar las probabilidades de transición, se requiere ocasionalmente una realización (ϕ, ψ) optimizada de la topología discreta que determinan la LTM's. La representación en coordenadas torsionales es generada para cuantificar el compromiso estructural de cada residuo, una cantidad fundamental para determinar el siguiente paso de plegamiento.

La estrategia computacional consiste básicamente en generar y almacenar iterativamente transiciones entre las cuencas atractivas de los mapas de Ramachandran de los residuos individuales. Dentro de esta representación, el tiempo es particionado en múltiplos del tiempo medio que requieren las transiciones entre cuencas para un residuo libre, el cual es un parámetro

del modelo [4]. La probabilidad de abandonar una cuenca depende de la profundidad de la misma, mientras que la probabilidad de llegar a una cuenca dada depende de su área lacunar. La ambigüedad estructural que surge de esta descripción módulo cuenca es eventualmente eliminada a través de la examinación de realizaciones geométricas optimizadas de la topología actual. Es más, esta geometría es necesaria para calcular la probabilidad P(k, t) del residuo k de cambiar de cuenca en el instante t, de acuerdo con el nivel de compromiso estructural del residuo en dicho instante de tiempo. El atractivo intuitivo de este método surge en dos aspectos: **(a)** grandes cambios conformacionales deberían ser probables sólo en aquellos residuos cuyo movimiento eventual tiene como consecuencia un cambio pequeño o negativo en la energía libre del sistema y **(b)** un algoritmo que genere caminos expeditivos de plegamiento no debería desperdiciar tiempo de cómputo buscando conformaciones en una región (ϕ, ψ) que ha sido previamente identificada como estructuralmente improductiva.

Una hipótesis central utilizada para diseñar el algoritmo es que las contribuciones locales a la energía intramolecular de una molécula pueden ser modeladas como un conjunto de restricciones geométricas locales variables en el tiempo que constituyen la LTM dependiente del tiempo. Estas restricciones definen regiones en el espacio torsional que son exploradas para obtener un patrón optimizado de interacciones de largo rango. Este mismo patrón determina el nuevo conjunto de restricciones y por lo tanto, dirige la generación del siguiente patrón. En otras palabras, en lugar de simular la evolución de la estructura, el algoritmo simula la evolución de restricciones torsionales codificadas como una LTM que varía a medida que el tiempo transcurre.

En cada iteración del algoritmo, se genera un patrón óptimo de interacciones de largo rango, compatible con el conjunto de restricciones torsionales definidos por la LTM actual. Este

patrón permite calcular el nivel de compromiso estructural de cada residuo de la cadena y, más concretamente, la probabilidad de cada residuo de cambiar la cuenca actual. Estas probabilidades son las que determinan la nueva LTM para la siguiente iteración.

Podemos reformular el esquema algorítmico presentado al final del capítulo 2 en término de los siguientes pasos, que representan una iteración genérica t del algoritmo:

I. Selección inicial de coordenadas (ϕ_i, ψ_i) para cada $i \in I(t)$, donde

I(t) = familia de residuos que cambiaron de cuenca en la iteración t-1

II. Optimización Monte-Carlo de estructuras imperfectas (secundarias y terciarias) que resultan de la selección de coordenadas en el paso **I** o bien que han sobrevivido iteraciones previas.

III. Para cada residuo k, se efectúa el siguiente cálculo:

$$\Delta U(k) = \sum_{(i,\, j) \in J(k)} U^{eff}(i, j)$$

donde J(k) = { pares (i, j) con $i \leq k \leq j$ y $U^{eff}(i, j) \neq 0$ }. Si k realizara un cambio conformacional importante (como lo es saltar de cuenca), las interacciones que se verían afectadas son justamente aquellas entre los pares (i, j) que forman J(k).

IV. Para cada residuo k se calcula:

$$\Delta S_{sc}(k) = \sum_{(i,\, j) \in J_C(k)} \Delta S_{sc}(i, j)$$

donde $\Delta S_{sc}(i, j)$ representa la pérdida entrópica de las cadenas laterales asociada con la formación del contacto (i, j), y $J_c(k)$ = { pares (i, j) con $i \leq k \leq j$ y $U^{eff}(i, j) < 0$ }.

V. Para cada residuo k contenido en una región de loop se calcula la pérdida entrópica conformacional asociada con formar dicho loop, utilizando para esto la expresión de Jacobson-Stockmayer:

$$\Delta S_{loop}(k) = R\,\lambda \ln(\,L(k)\,)$$

donde L(k) es la longitud del loop renormalizado que contiene a k y λ = 1,75. El loop renormalizado se construye eliminando las regiones del loop original que estén involucradas en motivos estructurales contenidos dentro de éste [4].

VI. Para cada residuo k se calcula el cambio en energía libre asociado con el eventual cambio de cuenca por parte de k:

$$\Delta G(k) = \Delta U(k) - T\,\{\,\Delta S_{sc}(k) + \Delta S_{loop}(k)\,\}$$

y en función de esta cantidad se calcula la probabilidad de cambiar de cuenca en el intervalo (t, t+1) con la siguiente expresión:

$$P(k) = \exp[\Delta G(k)/RT]$$

VII. Utilizando las probabilidades P(k) se determina el conjunto I(t+1), y para cada uno de los residuos en I(t+1) se seleccionan nuevas cuencas de acuerdo con las áreas lacunares.

El resultado de esta iteración es transformar un estado topológico de la cadena, LTM(t), en el siguiente, LTM(t+1), una transición que representa τ = 10 ps en tiempo real, de acuerdo con nuestro ajuste paramétrico de los datos experimentales [4]. El tiempo medio para una transición de cuencas en un residuo libre ha sido determinado correlacionando los eventos en nuestra simulación con la cronología experimental de los eventos de plegamiento más significativos, en particular, la aparición del núcleo colapsante en los plegadores de dos estados y el tiempo total de plegamiento [1, 2-3, 5, 6-12].

En las secciones subsiguientes, describiremos cada uno de los pasos indicados de manera esquemática recién.

4.2 Selección inicial de coordenadas

En este paso se efectúa una asignación de coordenadas (ϕ, ψ) dentro de las cuencas seleccionadas para LTM(t). La asignación es restringida a aquellos residuos que pertenecen a la familia I(t) formada por todos aquellas unidades que cambiaron sus cuencas en la iteración previa que definió la transición LTM(t-1) $\rightarrow$ LTM(t). Esta asignación realiza utilizando una distribución de probabilidades intracuenca que asigna distintos pesos a cada uno de los isómeros torsionales dentro de una misma cuenca. En la implementación actual del algoritmo hemos adoptado una distribución empírica para cada residuo obtenida de una base de datos estructural generada por el programa PROCHECK [13].

4.3 Optimización Monte-Carlo

Este paso consiste en un procedimiento de optimización cuya función es la de mejorar estructuras imperfectas que correspondan a una determinada LTM. Esta subrutina es activada solamente cuando se producen cambios de cuenca en la LTM, que puedan eventualmente definir nuevos patrones estructurales locales y/o interacciones de largo rango. Ya que una cuenca de Ramachandran presenta una gran latitud conformacional, es altamente improbable que una elección inicial no correlacionada de coordenadas torsionales dentro de un conjunto de cuencas determinado genere una estructura óptima compatible con las restricciones (ϕ, ψ) impuestas por la permanencia dentro de dichas cuencas.

Antes de la optimización propiamente dicha, un paso de reconocimiento de estructuras es necesario, acorde con la gramática de patrones LTM previamente introducida (sección 1.5 del

capítulo 1). En función de esta gramática, los motivos estructurales secundarios son reconocidos de acuerdo con las siguientes reglas:

(a) Una cadena de cinco o más residuos consecutivos en cuenca 2 es identificada como una hélice α dextrógira (levógira si tenemos 3's en lugar de 2's). El tamaño de la ventana es fundamental, ya que una ventana más pequeña formada por 2's puede indicar alternativamente un β-turn.

(b) Una cadena de dos o más 1's consecutivos es identificada como un potencial strand β, el cual es optimizado posteriormente sólo si forma parte de una hoja β.

(c) Diferentes cadenas pueden ser leídas como una región de β-turn, por ejemplo: 2223 o 1222. En general, cualquier cadena de cuatro cuencas consecutivas que estén produciendo un puente de hidrógeno (i, i+3) es interpretada como un turn.

Una vez que las regiones estructurales básicas han sido identificadas, el patrón de contactos es inspeccionado para encontrar contactos de largo rango entre las regiones detectadas. Esto permite localizar las hojas β y las interacciones terciarias entre hojas y hélices. La optimización es efectuada en tres etapas:

(1) Todas las estructuras imperfectas son identificadas. Por ejemplo, una hoja β es clasificada como imperfecta cuando sus strands no están bien alineados en un esquema regular (paralelo o anti-paralelo), o bien cuando presenta un número muy escaso de puentes de hidrógeno entre los strands. En el caso de las hélices α, el criterio de imperfección consiste solamente en que el número de puentes de hidrógeno a lo largo de la hélice sea menor que cierto valor crítico. Finalmente, para interacciones terciarias de tipo hoja-hélice, un interacción es reconocida como

imperfecta si no han sido saturadas todas las posibles interacciones atractivas entre las estructuras constituyentes.

(2) Para cada estructura clasificada como imperfecta en la etapa anterior, los siguientes conjuntos de residuos son construidos:

Ξ = Todos los residuos en la cadena que participan en las interacciones que estabilizan la estructura en cuestión. Por ejemplo, para una hoja β, este conjunto está formado por las regiones de strand.

Ω = Residuos que tienen la capacidad de mejorar las interacciones de los residuos en Ξ, vía la modificación de sus coordenadas torsionales. Tomando nuevamente como ejemplo el caso de la hoja b, Ω contiene solamente a los residuos del loop o β-turn intermedio.

(3) Para cada estructura imperfecta, la energía de las interacciones de la región Ξ:

$$U_{\Xi} = \sum_{i,j \in \Xi} U^{eff}(i, j)$$

es optimizada a través de una exploración Monte-Carlo de las conformaciones torsionales accesibles a los residuos de la cadena que pertenecen a la región Ω. Esta optimización es efectuada con dos restricciones: **(a)** la exploración torsional debe permanecer dentro de las cuencas asignadas (el optimizador no puede perturbar la dinámica de cuencas) y **(b)** la energía restante del sistema, $U^{eff} - U_{\Xi}$, no puede aumentar durante la búsqueda, si bien no es requerido que disminuya.

Esta manera de llevar a cabo la exploración intracuenca, por medio de una búsqueda limitada a mejorar hélices α, hojas β e interacciones α-β, resulta finalmente ser mucho más eficiente que una búsqueda exhaustiva del mínimo de la función potencial. Esto se debe no solamente a que una búsqueda exhaustiva consume una gran cantidad de tiempo computacional,

sino también porque una búsqueda detallada que utilice una representación aproximada de la proteína, en particular de las cadenas laterales, no conduce a mínimos realistas: una conformación que presente un empaquetamiento óptimo de cadenas laterales modeladas como elipsoides puede resultar muy diferente de las estructuras usualmente adoptadas por una proteína natural, y por lo tanto una excesiva optimización puede atrapar al sistema en plegamientos irreales.

4.4 Cálculo del compromiso estructural de un residuo

Como fue señalado antes, el compromiso estructural de un residuo k es esencial para calcular su probabilidad de cambiar de cuenca. Para calcular esta probabilidad, es necesario evaluar el efecto que tendría el eventual cambio de cuenca del residuo k en cuestión. Para esto definimos la familia J(k) de pares de residuos (i, j) tales que $i \leq k \leq j$ y $U^{eff}(i, j) \neq 0$. Esta es precisamente la familia de interacciones atractivas (contactos) o repulsivas (anticontactos) cuya supervivencia depende de la conformación torsional adoptada por el residuo k. La cantidad $\Delta U(k) = \sum_{(i, j)\in J(k)} U^{eff}(i, j)$ es la pérdida energética asociada con las interacciones que serían desmanteladas si k cambiase de cuenca. Notemos que $\Delta U(k)$ puede ser positivo, en el caso de que las interacciones desmanteladas fuesen anticontactos, por lo que en este caso el algoritmo debería favorecer el cambio de cuenca para la unidad k.

4.5 Cálculo de las entropías de las cadenas laterales

Debido a que utilizamos como espacio conformacional base a $X = T^{2N}$, nuestro modelo no incorpora explícitamente la geometría de las cadenas laterales. En cambio, el movimiento

torsional de las cadenas laterales esta integrado como una contribución entrópica que depende de la conformación del backbone. Cada vez que un contacto hidrofóbico es formado, se produce una pérdida entrópica de las cadenas laterales originada en la restricción conformacional que sufren las cadenas que forman parte del contacto. Sea z(k) el número de coordenadas diedrales en el residuo libre k. Cuando k forma un contacto que inmoviliza parcialmente al m-ésimo diedro ($m \leq z(k)$), podemos definir $s_{k,m}$ como la medida del perímetro de la porción del círculo unitario accesible a la variable torsional m cuando está formando parte del contacto. Como el "volumen torsional" total accesible a la variable m-ésima en la conformación libre es 2π, entonces podemos escribir el cambio entrópico total asociado al contacto (i, j) como:

$$\Delta S_{sc}(i, j) = R \{ \ln(s_{i,m(i, j)} / 2\pi) + \ln(s_{j,m(j, i)} / 2\pi) \} = \delta S_{sc}(i) + \delta S_{sc}(j)$$

donde m(i, j) es la variable diedral afectada en el residuo i por la formación del contacto (i, j). La expresión $\delta S_{sc}(i) = R \ln(s_{i,m(i, j)} / 2\pi)$ es válida cuando el número de contactos que involucran al residuo i al momento de evaluar el contacto (i, j) es menor que el número z(i), en caso contrario la cadena lateral i no puede perder más libertad conformacional y por lo tanto $\delta S_{sc}(i) = 0$. Estas definiciones operacionales tienen en cuenta el hecho de que la barrera entrópica para formar los primeros contactos que involucran a cierta unidad es muy alta, pero una vez que estos contactos iniciales son formados, los siguientes son más fáciles de concretar porque la cadena lateral ya está parcialmente inmovilizada.

Los valores z para cada aminoácido se obtuvieron simplemente contando el número de diedros (sin restricciones) que poseen las cadenas laterales cuando están libres, y son mostrados en la Tabla IId. La siguiente expresión ha sido adoptada para simplificar los cálculos:

$$2\pi / s_{i,m(i, j)} \approx c$$

donde c = 2,6 representa el factor de restricción torsional de las cadenas laterales. Entonces:

$$\delta S_{sc}(i) = - R \ln(c) \text{ si } n(i) \leq z(i)$$

$$\delta S_{sc}(i) = 0 \qquad \text{si } n(i) > z(i)$$

donde n(i) es el número de contactos hidrofóbicos que involucran al residuo i al momento de calcular $\delta S_{sc}(i)$.

De manera similar que para el cálculo de $\Delta U(k)$, se define la pérdida entrópica asociada con las interacciones atractivas (contactos) que serían desmanteladas durante el eventual cambio de cuenca de la unidad k como:

$$\Delta S_{sc}(k) = \sum_{(i,\, j) \in J_c(k)} \Delta S_{sc}(i, j) = \sum_{(i,\, j) \in J_c(k)} \{ \delta S_{sc}(i) + \delta S_{sc}(j) \}$$

donde $J_c(k)$ es la familia formada por los pares de residuos (i, j) que verifican $i \leq k \leq j$ y $U^{eff}(i, j) \neq 0$.

4.6 Cálculo de las entropías de los loops

Dado que una unidad k está ocupando una región de loop, hay una contribución entrópica $\Delta S_{loop}(k)$ que favorece que k cambie de cuenca: el desmantelamiento del loop trae aparejada una mayor libertad conformacional de la cadena. Para calcular la pérdida mínima en entropía conformacional del backbone, $\Delta S_{loop}(k)$, asociada con la clausura del loop que contiene a k, utilizamos la aproximación de Jacobson-Stockmayer:

$$\Delta S_{loop}(k) = R \lambda \ln L(k)$$

donde $\lambda = 1{,}75$ y L(k) es el número de unidades que forman al loop renormalizado que contiene a k. El loop renormalizado es construido identificando el contacto (i, j) que cierra el loop, y

luego removiendo del segmento [i, j] de la cadena a todos aquellos residuos que están involucrados en estructuras secundarias.

Esta contribución entrópica es necesaria para asegurar la cooperatividad en el surgimiento de organización de largo rango: si un contacto de largo rango es formado, su fragilidad se debe a que tiene un alto costo entrópico y por lo tanto, primero debe ocurrir una progresiva disminución del tamaño del loop a través de sucesivos pasos de renormalización.

Los residuos en un loop de gran tamaño tienen un término $-T\Delta S_{loop}$ muy positivo, lo cual favorecerá su cambio de cuenca. Este cambio provocará con seguridad la ruptura del contacto de largo rango que define al loop. Por otro lado, si el contacto de largo rango aparece solamente luego de que varios contactos se han formado dentro de su loop asociado, entonces la contribución entrópica $-T\Delta S_{loop}$ para cualquier residuo en el loop renormalizado será más pequeña que antes y, en consecuencia, estos residuos tendrán una tendencia menor a cambiar de cuenca. Esto implica que los contactos de largo rango formados de una manera jerárquica son más estables que aquellos formados con un gran costo entrópico, lo cual está en pleno acuerdo con observaciones experimentales.

4.7 Cálculo de las probabilidades de cambio de cuenca

Una vez que las cantidades $\Delta U(k)$, $\Delta S_{sc}(k)$ y $\Delta S_{loop}(k)$ asociadas al eventual cambio de cuenca de la unidad k han sido calculadas, podemos evaluar el cambio en energía libre (medida termodinámica del compromiso estructural del residuo) que ocasionaría dicho movimiento. Esta cantidad termodinámica está dada, para el residuo k, por la siguiente expresión:

$$\Delta G(k) = \Delta U(k) - T\Delta S(k) = \Delta U(k) - T\ \{\ \Delta S_{loop}(k) + \Delta S_{sc}(k)\ \}$$

Entonces la probabilidad de que el residuo k cambie de cuenca está determinada por:

$$P(k) = \exp[\Delta G(k) /RT] = \exp[\{ \Delta U(k) - T \{ \Delta S_{sc}(k) + \Delta S_{loop}(k) \} \}/ RT]$$

cuando $\Delta G(k) < 0$, caso contrario fijamos $P(k) = 1$. Es decir que la probabilidad del residuo de cambiar de cuenca depende en la estabilidad del patrón que se pierde como consecuencia de la transición de cuenca. Esto refleja el hecho de que la profundidad (o equivalentemente, la barrera de activación para el pasaje de una cuenca a otra) de una cuenca de Ramachandran aumenta a medida de que el residuo se vuelve más comprometido estructuralmente.

4.8 Generación de I(t+1) y LTM(t+1)

Un residuo k está contenido en el conjunto I(t+1) si un número aleatorio r elegido uniformemente en el intervalo [0, 1] verifica $r < P(k)$. Para tales residuos, una nueva cuenca es seleccionada de acuerdo con las siguientes probabilidades:

$$P(\text{seleccionar cuenca } y_k) = A(y_i) / \sum_{y''_k \neq y'_k} A(y''_i) \quad \text{si } y_k \neq y'_k$$

$$= 0 \qquad \text{si } y_k = y'_k$$

donde y'_k es la cuenca ocupada por el residuo k durante la iteración t y $A(y_k)$ es el área lacunar de la cuenca y_k.

De esta manera, la nueva conformación topológica LTM(t+1) es generada fácilmente. El lector debería notar que, a diferencia de la probabilidad de cambiar de cuenca, la probabilidad de seleccionar una cuenca es independiente de su profundidad y depende solamente de su área lacunar.

4.9 Referencias

[1] Prediction of protein folding mechanisms from free energy landscapes derived from native structures. E. Alm, D. Baker. *Proc. Natl. Acad. Sci. USA* **96**, 11305-11310 (1999).

[2] A simple model for calculating the kinetics of protein folding from three dimensional structures. V. Muñoz, W. A. Eaton. *Proc. Natl. Acad. Sci. USA* **96**, 11311-11316 (1999).

[3] Three key residues from a critical contact network in a protein folding transition state. M. Vendruscolo, E. Paci, C. Dobson, M. Karplus. *Nature* **409**, 641-645 (2001).

[4] Conformational-dependent environments in folding proteins. A. Fernández. *J. Chem. Phys.* **114**, 2489-2502 (2001).

[5] The barriers in protein folding. T. R. Sosnick, L. Mayne, R. Hiller, S. W. Englander. *Nature Struct. Biol.* **1**, 149-156 (1994).

[6] D/H amide isotope effects reveal when hydrogen bonds form during protein folding. B. Krantz, L. Moran, A. Kentsis, T. R. Sosnick. *Nature Struct. Biol.* **7**, 62-71 (2000).

[7] The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the protein folding nucleus. Z. Guo, D. Thirumalai. *Fold. Des.* **2**, 377-391 (1997).

[8] Specific nucleus as the transition state for protein folding: evidence from the lattice model. V. I. Abkevich, A. M. Gutin, E. I. Shakhnovich. *Biochemistry* **33**, 10026-10036 (1994).

[9] Finding the collapse-inducing nucleus in a folding protein. A. Fernández, G. Appignanesi, A. Colubri. *J. Chem. Phys* **114**, 8678-8684 (2001).

[10] Molecular collapse: the rate-limiting step in two-state cytochrome c folding. T. R. Sosnick, L. Mayne, S. W. Englander. *Proteins* **24**, 413-426 (1996).

[11] Folding and stability of a tryptophan-containing mutant of ubiquitin. S. Khorasanizadeh, I. D. Peters, T. R. Butt, H. Roder. *Biochemistry* **32**, 7054-7063 (1993).

[12] Structural and kinetic characterization of early folding events in b-lactoglobulin. K. Kwata, et al. *Nature Str. Biol.* **8**, 151-155 (2001).

[13] Main-chain bonds lengths and bond angles in protein structures. R. A. Laskowski, D. S. Moss, J. M. Thornton. *J. Mol. Biol.* **231**, 1049-1067 (1993).

Capítulo 5

Resultados y predicciones

5.1 Resultados teóricos: capturando la dinámica esencial que subordina el proceso de plegamiento

Desde el punto de vista teórico, esta tesis nos ha conducido a dos resultados centrales, uno de ellos de naturaleza matemática y general, y el otro inscripto específicamente en el problema del plegamiento de proteínas.

En el capítulo 2 introducimos un formalismo matemático que, hasta donde sabemos, es novedoso en el campo de los procesos estocásticos y que estudia dos conceptos relacionados: la proyección de procesos markovianos y la conmutatividad de dinámicas markovianas. La idea física que motiva estos conceptos es la de describir una dinámica detallada a través de una descripción "grosera" o discretizada del espacio de conformaciones sin comprometer o perder información estructural que podría afectar la dinámica. Los resultados obtenidos son interesantes por si mismos y además tienen una interpretación física intuitiva. Ya que el estudio de sistemas biológicos complejos es de gran importancia en la actualidad (y tendrá más importancia en el futuro), un modelo matemático que permita un tratamiento simplificado, pero a la vez riguroso, de estos sistemas podría representar una herramienta de gran utilidad conceptual y predictiva.

Hay que indicar que esta teoría matemática se encuentra en su etapa inicial, sólo se presentaron las definiciones fundamentales y se dedujeron algunos resultados elementales que fueron de utilidad inmediata en el problema del plegamiento. Es posible que en un futuro

cercano se puedan obtener nuevos resultados y conclusiones de mayor sofisticación matemática, lo cual hace a esta teoría particularmente atractiva.

La aplicación de este formalismo matemático en el problema particular del plegamiento nos permitió construir una dinámica discreta o "digital" en el espacio de cuencas de Ramachandran: la dinámica "módulo cuencas". La validez de este modelo matemático sustenta el algoritmo de plegamiento presentado en la tesis. La gran ventaja de este algoritmo es que, al no depender de todos los detalles conformacionales de una proteína, es capaz de reproducir la expeditividad y robustez del plegamiento. Además, y al contrario de lo que sucede con los tratamientos "clásicos" tales como dinámica molecular, el algoritmo permite alcanzar las escalas de tiempo biológicamente relevantes a efectos de predecir las estructuras nativas de proteínas naturales (de 1 milisegundo a algunos segundos). En la sección siguiente nos dedicaremos precisamente a analizar algunos resultados generados con el algoritmo y a dar una idea de su poder predictivo.

5.2 Resultados computacionales: ¿De cuántos caminos de plegamiento dispone una proteína? ¿Que propiedad determina la diversidad de caminos?

Como se anticipó al comienzo de la tesis, a efectos de testear el algoritmo de plegamiento hemos seleccionado dos proteínas de dominio simple: ubiquitina (1ubi) y la variante hipertermófila de la proteína G (1gb4), las cuales se cree que son plegadores de dos estados. Ambas proteínas tienen una estructura espacial similar, sin embargo, las simulaciones muestran que dependen de manera muy diferente en los contextos de largo rango para encontrar sus respectivos plegamientos nativos.

Comprobamos que estas proteínas se pliegan siguiendo el modelo cinético de dos estados, y encontramos una característica que suponemos debe ser genérica a cualquier plegador de dos estados: una disminución en la amplitud de las fluctuaciones estructurales es alcanzada solamente cuando el número de puentes de hidrógeno altamente protegidos alcanza un régimen constante o "plateau" que corresponde a un máximo estacionario. Esto nos permite identificar el núcleo colapsante y mostrar que el plegamiento no se vuelve expeditivo hasta el momento en que se genera una topología capaz de proteger del ataque del agua a los puentes de hidrógeno intramoleculares. Asimismo, generamos las trayectorias individuales más representativas que forman los ensambles del estado de transición de 1ubi y 1gb4.

5.2.1 Identificando el núcleo conducente del plegamiento de una proteína

En primer lugar, examinamos las características y patrones dependientes del tiempo que se obtienen a partir los caminos de plegamiento generados con nuestro algoritmo. Los resultados sostienen el escenario cinético de dos estados propuesto para proteínas de dominio simple con longitud < 120 [1-7], en contraste con un modelo cinético de tres estados propuesto para 1ubi en [8], y además nos permiten caracterizar estructuralmente el núcleo del plegamiento a lo largo de cada camino específico.

Cada una de las corridas está formada por 10^6 iteraciones y una conformación es identificada como estacionaria si sobrevive más allá de 10^5 iteraciones (1 ms). El estado inicial en todas las corridas se obtiene por una atribución aleatoria de cuencas de Ramachandran para cada residuo, en concordancia a la distribución basada en sus áreas lacunares relativas.

Una corrida típica altamente reproducible para 1gb4 a T = 323 K es descripta en Figs. 7 y 8. Fig. 7a revela una disminución dramática en las fluctuaciones estructurales que ocurre luego

de 1,6 ms. Esta disminución está marcada por un súbito descenso en el número total de residuos que cambian de cuenca en el instante t, $\Lambda = \Lambda(t)$. Más aún, hay una fuerte correlación entre la disminución de las fluctuaciones estructurales (Fig. 7a) y el arribo a un "plateau" cuasi-estacionario en el número V(t) de puentes de hidrógeno altamente protegidos (Fig. 7b). Un puente de hidrógeno se dice altamente protegido cuando se cumplen dos condiciones: **(a)** está rodeado por cinco o más residuos hidrofóbicos ubicados a una distancia menor de 7 Å con respecto a cualquiera de los dos residuos que forman el puente, y **(b)** tiene una distorsión angular menor que 37° grados. Estos resultados corresponden a una corrida exitosa y altamente reproducible, la cual, luego de 10^6 iteraciones (instantáneas de la simulación son mostradas en Figs. 8a-d), conduce a una estructura estacionaria muy similar al plegamiento nativo, esto es, a una distancia de Hamming de 3% con respecto a la matriz de distancias (Distances Matrix, DM, ver figura 4) de la estructura nativa. Reportamos los resultados siguiendo esta convención: entradas de color negro en la DM indican distancias entre carbonos α menores que 7,5Å, mientras que entradas grises indican distancias entre 7,5 Å y 8,5 Å.

Una comparación directa entre Figs. 7a y 7b sugiere que el núcleo que induce el colapso hidrofóbico es alcanzado solamente cuando una topología capaz de proteger un número máximo de puentes de hidrógeno es generada. Es por esto que, para dirigir al plegamiento más allá del régimen de prueba y error y avanzar constructivamente hacia el plegamiento nativo, la cadena debe actuar como un organizador endógeno y altamente efectivo del solvente con el objeto de inhibir el ataque simultáneo del agua sobre toda la estructura secundaria incipiente.

En corridas típicas para 1gb4, la estructura del núcleo puede ser determinada a través del examen de la conformación presente al momento de la repentina disminución en $\Lambda(t)$, o bien en

V(t). El núcleo o estado de transición para este camino es consolidado aproximadamente en 1,6 ms (Fig. 8b). La hélice nativa está presente en el núcleo, mientras que la hoja β paralela que une ambos extremos (Fig. 8d), está ausente. Es más, la hélice está protegida en uno de sus extremos por una distorsión en la hoja β anti-paralela nativa que se encuentra al final de la cadena (comparar la DM estacionaria objetivo de Fig. 8d con la DM nuclear de Fig. 8b). Esta hélice encuentra aún más protección en el otro extremo por medio de interacciones terciarias que establece con la hoja β nativa del inicio de la cadena. Estas interacciones terciarias están ausentes en el plegamiento nativo, al igual que la distorsión en la hoja β inicial. Por lo tanto, la cadena utiliza productivamente conformaciones transitorias no nativas en la débil hoja β inicial para proteger su estructura nativa nuclear, y el núcleo posee un patrón de protección de estructuras nativas que difiere del patrón nativo predecido (y real). Una examen directo de la estructura del núcleo (Fig. 8b) muestra que ninguna parte de la hélice carece de algún tipo de protección estructural terciaria, originada en el contexto de largo rango en el cual residuos hidrofóbicos distantes se aproximan a los puentes de hidrógeno amida-carbonilo. Es por esto que la hélice puede ser considerada como un elemento nuclear central y por lo tanto es esperable un considerable efecto isotópico cinético para esta proteína en experimentos de intercambio de protón-deuterio [1].

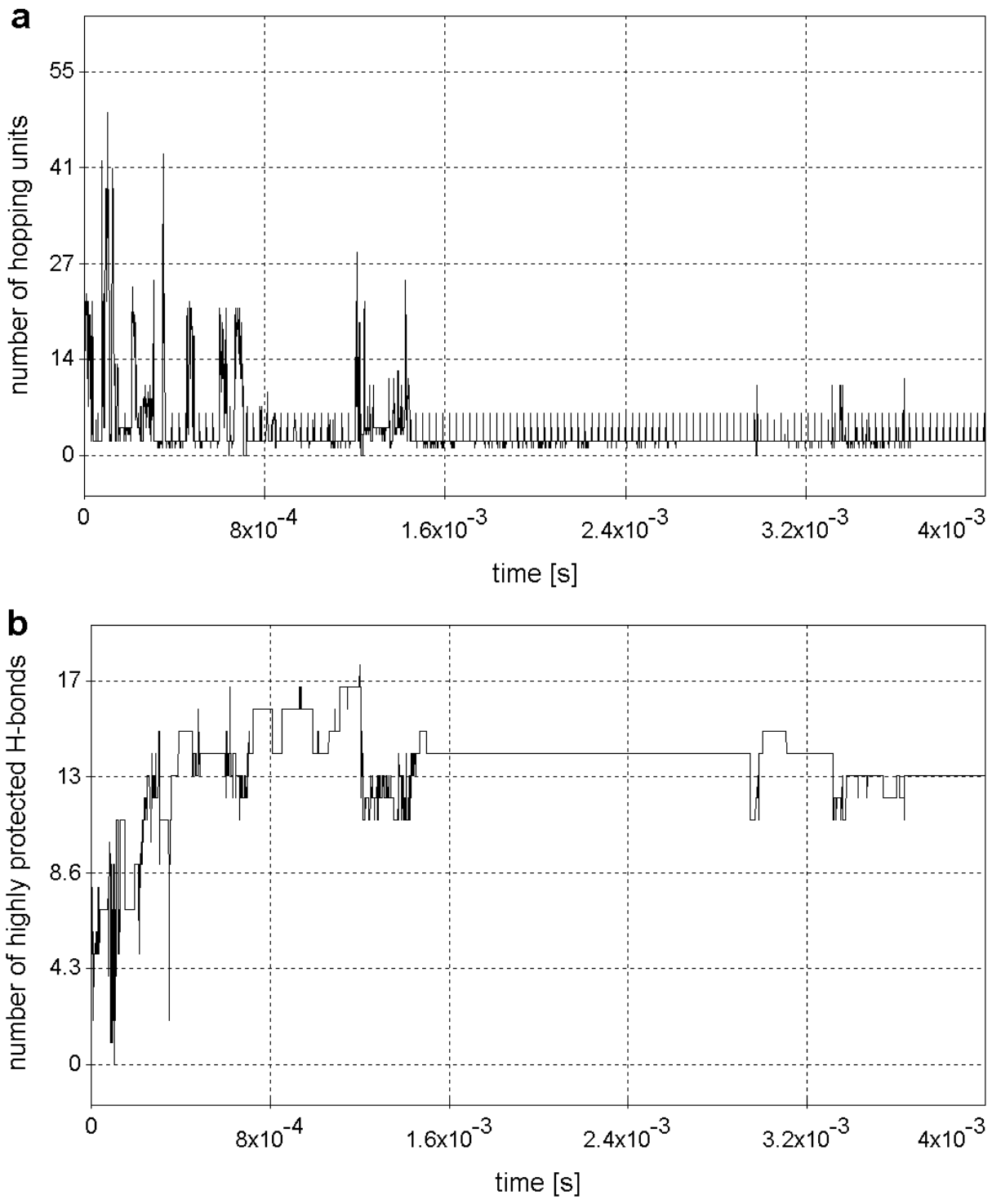


Figura 7: Dependencia en el tiempo del número de unidades que cambian de cuenca, $\Lambda = \Lambda(t)$, (a) y el número de puentes de hidrógeno altamente protegidos, V = V(t), (b), correspondientes a la simulación más probable para 1gb4 a T = 323 K.

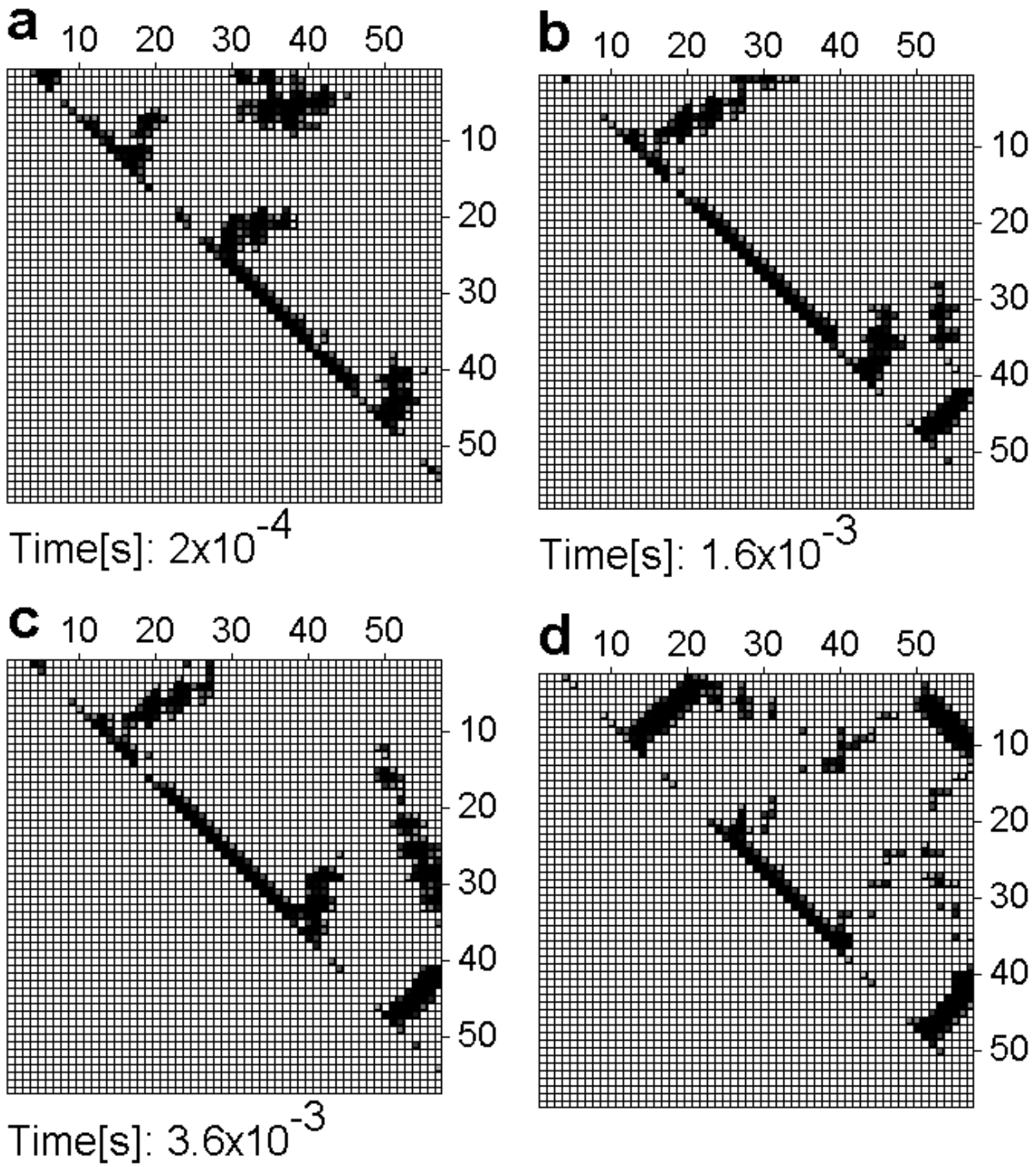


Figura 8: Instantáneas de la evolución de la DM a lo largo de la trayectoria más probable. Las tres primeras instantáneas (a, b, c) fueron tomadas en los instantes indicados en las figuras, mientras que (d) fue obtenida luego de completar la simulación (10^6 iteraciones).

Un camino altamente reproducible para 1gb4 está caracterizado en consecuencia por las siguientes características:

(a) Un núcleo que induce el colapso ocurre en 1,6±0,2 ms. Luego de este punto crítico, las fluctuaciones estructurales son frenadas y por lo tanto la formación del núcleo señala la superación de la única barrera importante en la energía potencial.

(b) El núcleo contiene alrededor de la mitad de la estructura secundaria nativa (Fig. 8b), con excepción de la hoja β paralela que une los extremos de la molécula (Fig. 8d) y la hoja anti-paralela inicial, que se encuentra severamente distorsionada (Fig. 8b). Todos los residuos de la hélice α encuentran protección a través de interacciones terciarias.

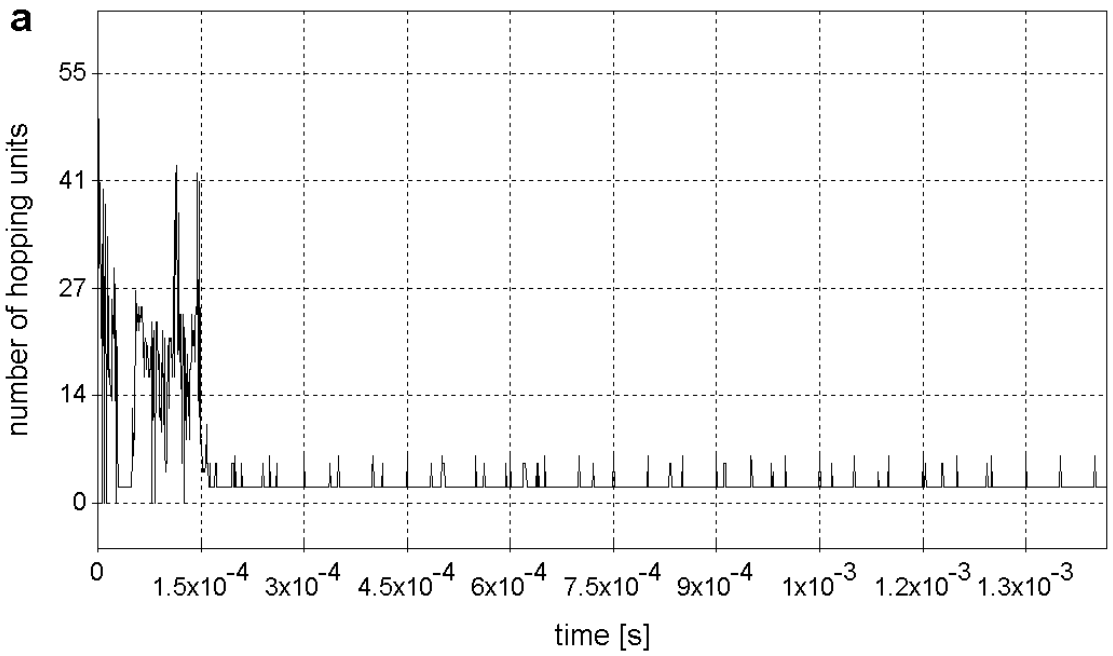
(c) Existe un cambio en los roles protectivos de los residuos hidrofóbicos a lo largo del plegamiento a medida que nos desplazamos desde la estructura nuclear hacia la estructura nativa (Fig. 8).

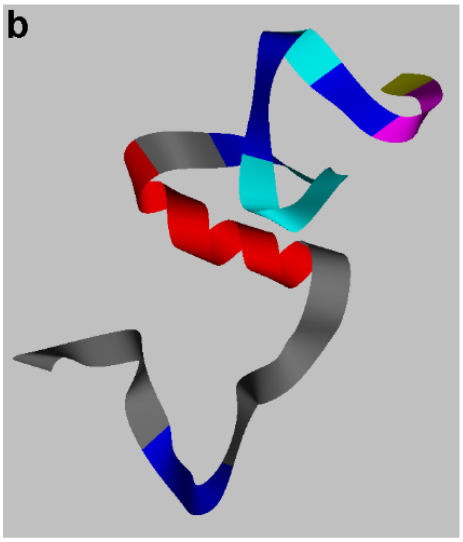
(d) Antes de la formación del núcleo, un período de "calma" puede ocurrir en $\Lambda(t)$ en corridas individuales, pero dichos períodos, tal como la ventana de tiempo 0,8-1,4 ms muestra en la Fig. 7a, no se superponen reproduciblemente en diferentes corridas. Incluso dentro de una sola corrida, ellos no constituyen períodos de verdadera calma, como es evidenciado por el comportamiento fluctuante de V(t) dentro de la misma ventana (Fig. 7b). Por lo tanto, no hay una acumulación significativa de intermediarios durante el plegamiento de esta proteína.

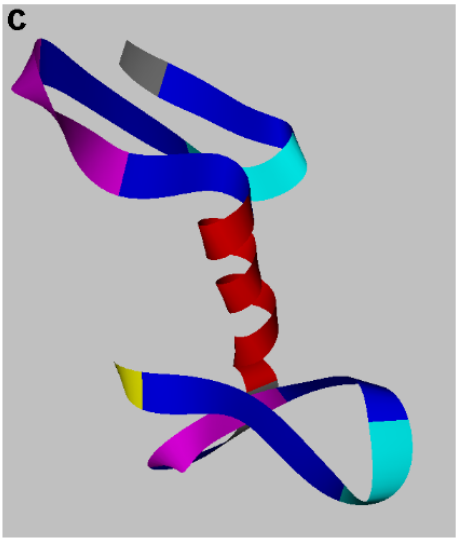
5.2.2 Mejorando la eficiencia del plegamiento: hacia una estrategia de diseño de proteínas

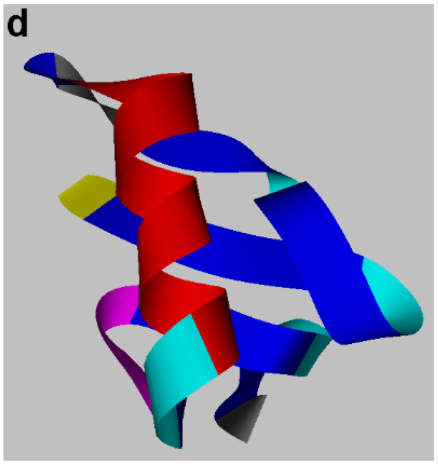
El examen de los patrones protectivos (Figs. 7, 8) para los puentes de hidrógeno del backbone, accesibles a través de nuestras simulaciones, nos permite el diseño de variantes mutacionales que mejoren la habilidad de la molécula para generar su núcleo protectivo. Esto puede ser ilustrado para 1gb4. La hoja β inicial de esta proteína es bastante anómala debido a que sus strands tienen una complementaridad hidrofóbica muy pobre, lo cual hace que esta hoja sea relativamente inestable [9]. En contraste con [9], nuestra estrategia básica para diseñar un plegador más expeditivo consiste en generar una sola mutación en el strand β que no sólo mejore la complementaridad de los strands, sino que también refuerce el patrón protectivo de la hélice α y, por lo tanto, la estabilidad del núcleo. Las Figs. 8 a-d muestran que una distorsión en la hoja β inicial es necesaria para proteger el extremo final de la hélice.

Figura 9 (siguiente página): Dependencia en el tiempo del número de unidades que cambian de cuenca ($\Lambda(t)$) (a) y tres instantáneas de la evolución del backbone para la 1gb4 mutante (LYS14LEU) a lo largo de su camino de plegamiento más probable (b, c, d).


a
number of hopping units
55
41
27
14
0
0
1.5x10-4
3x10-4
4.5x10-4
6x10-4
7.5x10-4
9x10-4
1x10-3
1.2x10-3
1.3x10-3
time [s]


b


c


d

En vista de lo anterior, resulta que introduciendo la mutación LYS14LEU obtendríamos mayor estabilidad en la hoja a través de los contactos ILE7-LEU14 e ILE8-LEU14. Por otra parte, la distorsión de la hoja β ya no sería necesaria para la estabilización del núcleo, debido a que LEU14 puede ser involucrado en la protección del puente de hidrógeno PHE31-ALA35, que participa en la nucleación de la hélice. Recíprocamente, PHE31 puede proteger los puentes de hidrógeno ínter-strand LEU14-ILE7 y LEU14-ILE8. Esta reciprocidad en las protecciones de los puentes de hidrógeno entre la hoja β inicial y la hélice α tiene como consecuencia que el núcleo sea mucho más fácil de formar en la variante mutada, tal como lo demuestra el dramático incremento en un orden de magnitud de la velocidad de plegamiento. Esto se evidencia en la temprana disminución de las fluctuaciones estructurales (Fig. 9a) y en las tres imágenes de la evolución del backbone mostradas en Fig. 9b-d que corresponden respectivamente en 10^{-4}, $1{,}4 \times 10^{-4}$ y $1{,}6 \times 10^{-4}$ s. La última conformación (Fig. 9d) es topológicamente indistinguible del plegamiento nativo y fue alcanzada precisamente en $t = t^* = 1{,}6 \times 10^{-4}$ s.

5.2.3 Generando caminos de plegamiento para la ubiquitina

Para ilustrar la versatilidad del algoritmo y su poder predictivo en diversos escenarios, fueron efectuadas corridas para 1ubi con el objeto de determinar su núcleo de plegamiento en caminos individuales. Para ahorrar tiempo de cómputo, la mayor parte de las corridas (de 10^6 iteraciones cada una) fueron realizadas secuencialmente: la cadena explora su espacio de conformaciones a medida que crece con la acreción secuencial de aminoácidos. Los residuos son agregados a una tasa constante comparable con las velocidades ribosomales de síntesis, de tal manera que la cadena completa de 1ubi fue ensamblada en 3,3 ms. Las simulaciones

secuenciales y no secuenciales dieron el mismo tiempo medio de colapso t* = 5,4 ms y las mismas estructuras para el núcleo. Tales resultados son consistentes con la cinética de plegamiento expuesta antes: el descenso en las fluctuaciones estructurales demanda la formación de una topología auto-protectiva como resultado de un período de prueba y error con extensas fluctuaciones que involucran hasta un 70% de los residuos (Fig. 10).

Una corrida altamente reproducible para 1ubi a T = 312K (correspondiente a un 32% de las corridas exitosas) es reproducida en Fig. 10 a través de un histograma LTM, y en Figs. 12a-c, d-f, a través de una secuencia de tres DM's significativas en la vecindad de la región crítica $t \approx t^*$ y sus backbones asociados. La cuenca de Ramachandran ocupada por cada residuo en cada instante es especificada en Fig. 10. El color azul representa cuenca 1; rojo, cuenca 2; verde, cuenca 3 y gris, cuenca 4 (solo presente en glicina). Un indicador del éxito de la simulación es el hecho de que la LTM estacionaria es idéntica a la correspondiente al plegamiento nativo (mostrada en la parte superior del histograma). Además, la DM estacionaria asociada difiere en menos del 1% de la DM nativa (Fig. 12c). Nuevamente, una inspección directa de los comportamientos de los gráficos de $\Lambda(t)$ y V(t) en una vecindad de t* (Fig. 11a, b) revela un mecanismo de nucleación basado en la formación de una topología auto-protectiva evidenciada por la maximización de puentes de hidrógeno altamente protegidos en t* y por un "plateau" para V(t) a partir de t*.

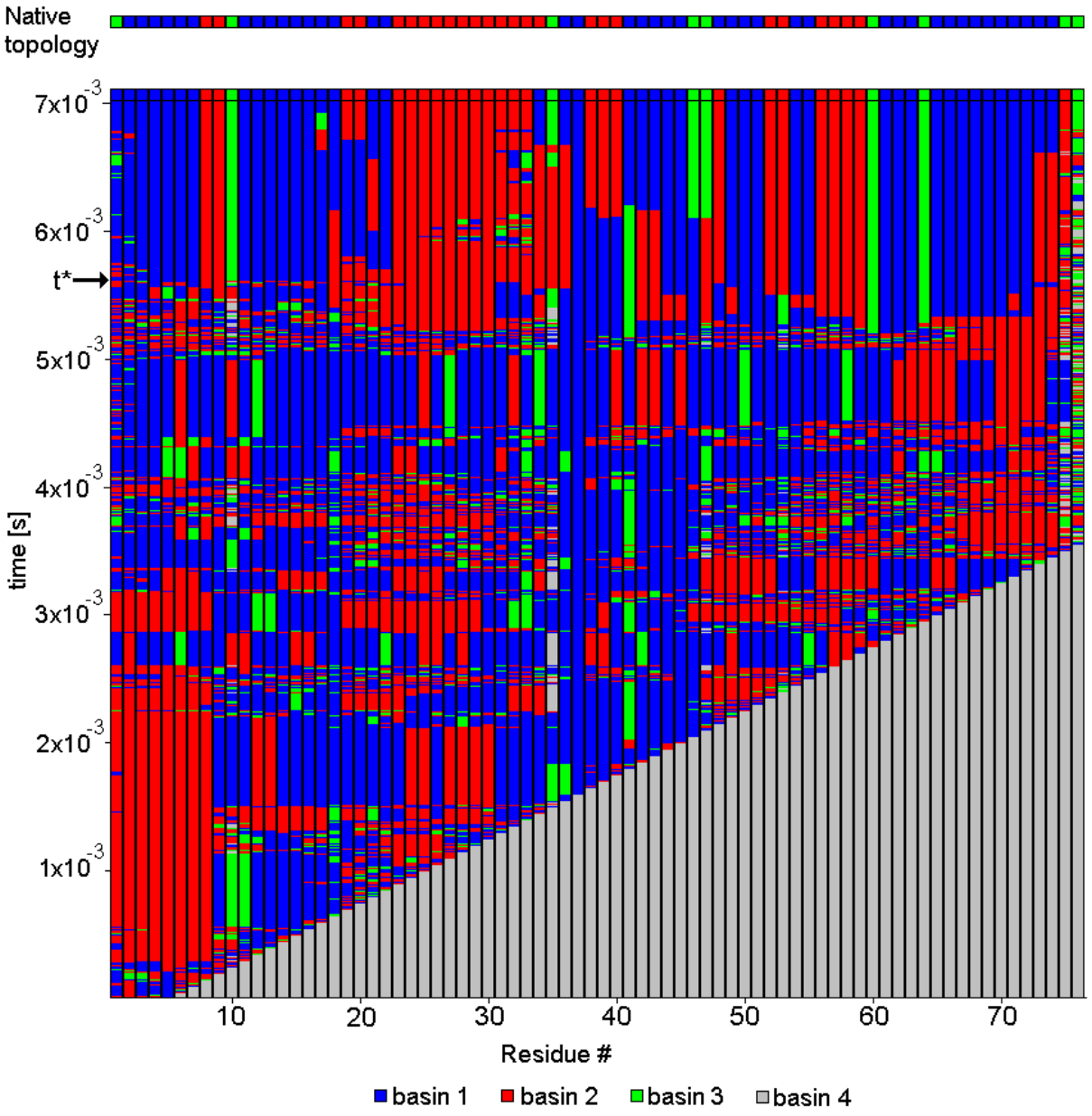


Figura 10: Histograma LTM para la simulación secuencial más probable del plegamiento de 1ubi a T = 313 K.

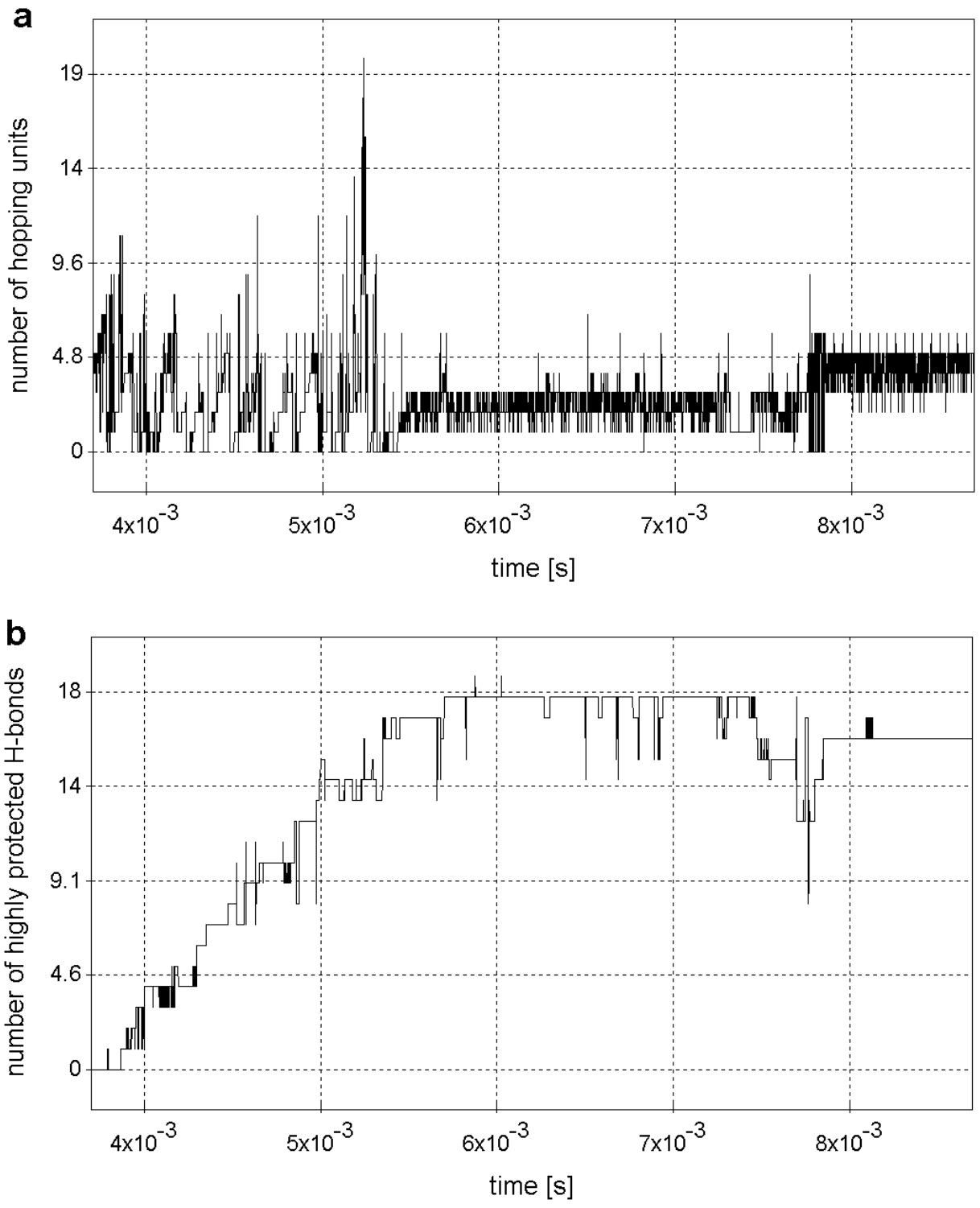


Figura 11 (siguiente página): Dependencia temporal del número de unidades que cambian de cuenca ($\Lambda(t)$) (a) y el número de puentes de hidrógeno altamente protegidos (V(t)) (b), correspondientes a la simulación más reproducible de 1ubi.

5.2.4 Disección computacional de la diversidad de caminos

En esta sección cuantificamos el nivel de diversidad de caminos, resolvemos el ensamble TS de acuerdo con familias de caminos de plegamiento y desarrollamos métodos para calcular promedios sobre ensambles TS para 1gb4 y 1ubi. En ambos casos, 72 corridas consistentes en 10^6 iteraciones cada una fueron seleccionadas de grupos de 132 y 155 simulaciones efectuadas a T = 323K y T = 313K, respectivamente. El grado de éxito de las corridas fue medido con los estándares descriptos previamente.

Nuestros resultados revelan que cada componente del ensamble TS de ambas proteínas está constituido por estructuras que comparten una topología común (LTM) pero que presentan diferentes DM's. Se observan grandes diferencias en los tiempos de formación de los núcleos cuando se comparan las diferentes rutas de plegamiento. El conjunto de LTM's en el ensamble TS es determinado detectando la LTM(t*) en los histogramas tales como los indicados en Fig. 10. A diferencia de la sección anterior, no se utilizó el plegamiento secuencial de 1ubi en este estudio. El peso estadístico P(M) del componente M del TS con LTM = LTM(t*) en el TS puede ser determinado contando el número de corridas que siguen rutas de plegamiento que pasan por la LTM(t*) en el instante t* y dividiendo por el número total de corridas generadas. Esta estimación empírica de P(M) coincide dentro de un error del 5% con la estimación teórica basada en el hecho de que la topología correcta (auto-protectiva) es el elemento requerido para dirigir al proceso de plegamiento hacia un régimen cuesta abajo a lo largo de la superficie de energía potencial monomolecular. Basados en esta premisa, podemos estimar P(M) analíticamente como:

$$P(M) = [\exp(-U(M)/RT)]\ [\prod_{j=1\ldots N} A_M(j)]\ / Z_{TS} \qquad (1)$$

donde U(M) es la energía potencial efectiva U^{ff} promediada sobre todas las conformaciones correspondientes a los caminos de plegamiento que en t* coinciden con la LTM(t*) del componente M, y donde la función de partición Z_{TS} es:

$$Z_{TS} = \sum_{M' \in TS} \exp(-U(M')/RT) \prod_{j'=1...N} A_{M'}(j') \qquad (2)$$

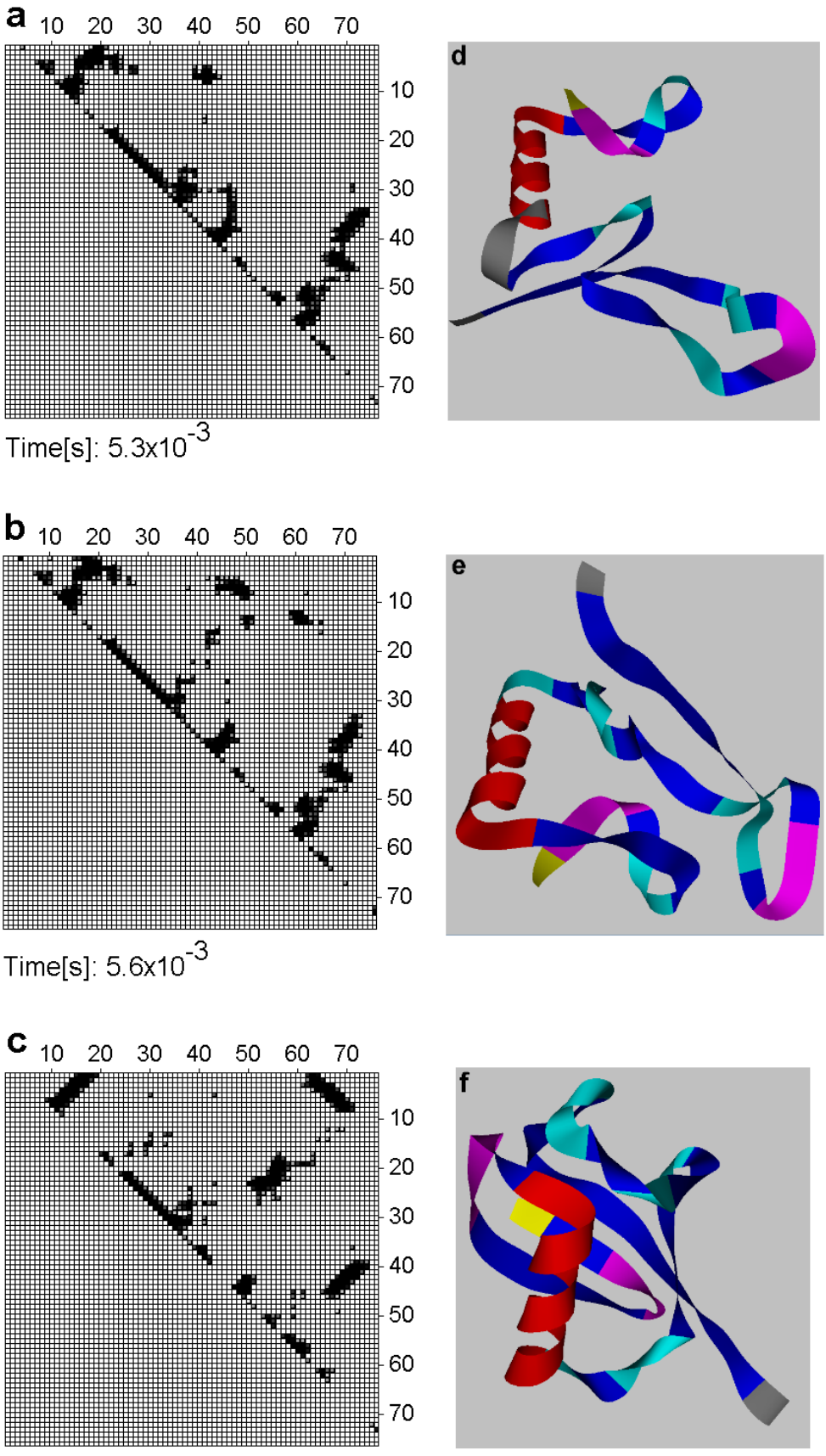
con $A_M(j)$ = área lacunar de la cuenca de Ramachandran del residuo j asignada por la LTM de la componente M en el instante t* = t*(M).

Los pesos estadísticos y la composición de los ensambles TS para 1gb4 y 1ubi son mostrados en Figs. 13a y b, respectivamente. Cada barra representa una componente (M) del ensamble TS caracterizado por una topología única. El ancho de la barra indica la cuantas DM's comparten la misma topología, cantidad que se calcula como función de su distancia de Hamming máxima. La altura de la barra indica el peso estadístico de la componente (Eqs. 1, 2).

Las Figuras 8b y 14 muestran la DM de los representantes generados más frecuentemente para cada componente (M = 1, 2, 3) en el ensamble TS de la 1gb4, identificado con el método indicado en la sección anterior. El tiempo t* en el cual cada estructura ocurre es indicado juntamente con el gráfico de la DM.

De manera similar para 1ubi, las DM's representativas asociadas con las LTM's en el tiempo crítico t* = t*(M) para los cinco componentes de su ensamble TS son mostradas en Figs. 12b y 15 para el componente principal (M = 1) y para los otros componentes (M = 2,..., 5), respectivamente.

Figura 12 (siguiente página): Tres instantáneas (a, b, c) de la evolución temporal de la DM para 1ubi a lo largo de la trayectoria más probable, tomadas en una vecindad de la región crítica t = t*. Los instantes en los que fueron tomadas las instantáneas son indicados en las figuras, excepto por (c), que corresponde a la estructura final (luego de 10^6 pasos de simulación). Las correspondientes conformaciones del backbone son mostradas en (d, e, f), respectivamente.


a
10 20 30 40 50 60 70
10
20
30
40
50
60
70
Time[s]: 5.3x10-3
d
b
10 20 30 40 50 60 70
10
20
30
40
50
60
70
Time[s]: 5.6x10-3
e
c
10 20 30 40 50 60 70
10
20
30
40
50
60
70
f

El análisis de los ensambles TS revela algunas características básicas que no serían reproducibles si la búsqueda conformacional estuviese dirigida por un potencial Gô [10]: la proteína es capaz de utilizar productivamente las conformaciones distorsionadas o no nativas en el transcurso de una búsqueda que no necesariamente representa un progreso monotónico hacia el plegamiento nativo. Por lo tanto, características estructurales secundarias no nativas aparecen en las tres componentes del ensamble TS de 1gb4: un patrón no nativo de protección de la hélice y un β-turn distorsionado para M = 1 (Fig. 8b), un patrón anti-paralelo junto con la hoja β paralela en M = 2 (Fig. 14a) y una hélice no nativa cerca del extremo final de la cadena en lugar de la hoja β paralela usual en M = 3 (Fig. 14b). En todos estos casos, características no nativas o distorsiones de las estructuras nativas son necesarias para proteger las estructuras secundarias nativas e imperfectas que nuclean el desarrollo estructural subsecuente.

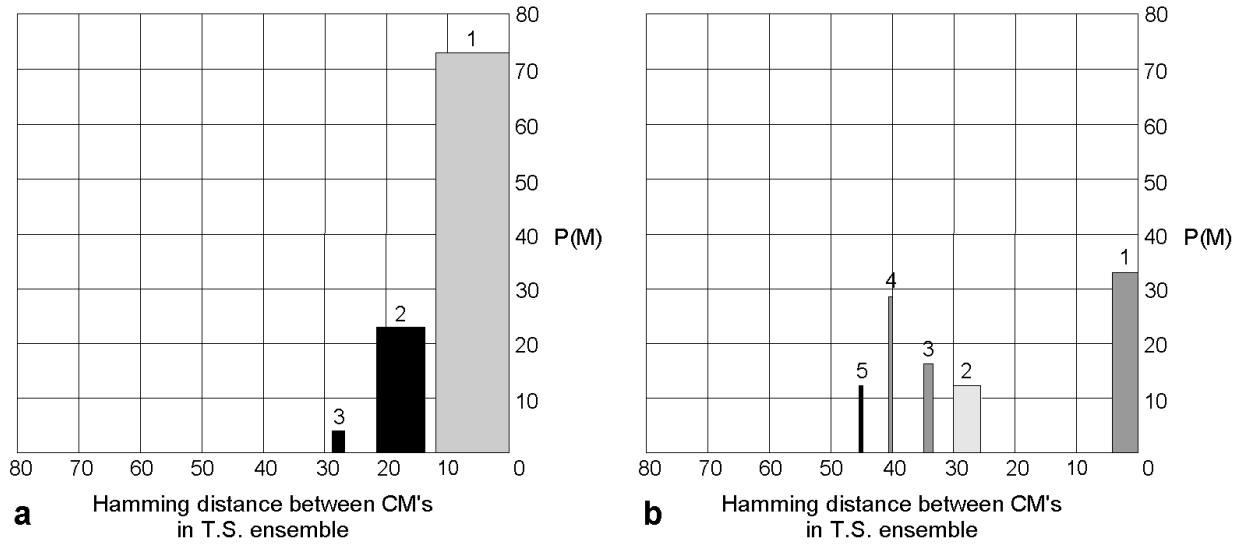


Figura 13: Composición del ensamble TS para 1gb4 (a) y 1ubi (b). Las topologías nucleantes definen componentes M = 1, 2,... que son indicadas por barras. La altura de cada barra representa su peso estadístico, mientras que su ancho señala la distancia de Hamming máxima entre las distintas DM's que comparten dicha topología.

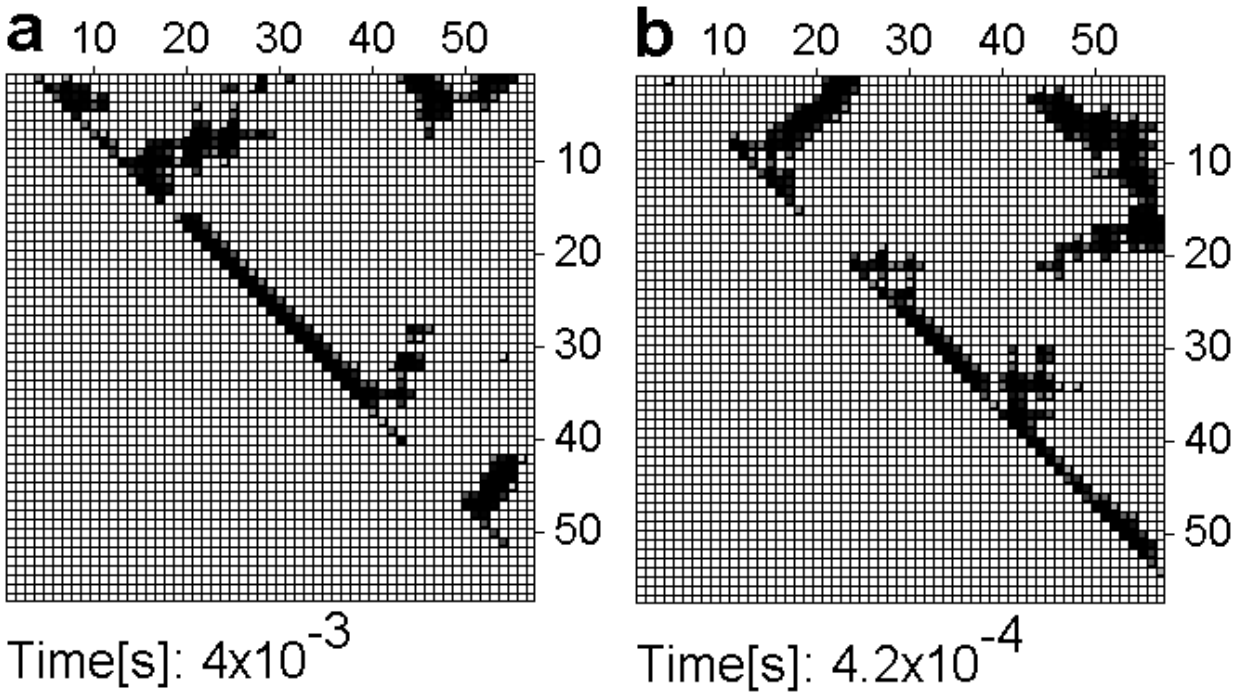


Figura 14: DM's representativas más frecuentes para las componentes M = 2 (a) y M = 3 (b) del ensamble TS de 1gb4.

Además de la ausencia de características nativas específicas en el núcleo, la ocurrencia de características secundarias no nativas y distorsionadas en el núcleo de plegamiento de 1ubi es también detectable: la componente M = 1 (Fig. 12b) muestra una interacción terciaria no nativa necesaria para apuntalar una hoja β nativa en el extremo inicial de la cadena; en M = 2 (Fig. 15a), la hélice α nativa 22-35 es ausente por completo mientras que la hélice α no nativa 55-65 está fuertemente protegida por la hoja β paralela nativa; M = 3 (Fig. 15b) revela nuevamente una hélice no nativa en la región 55-65, la hélice nativa 22-35 con un alto índice de distorsión y una hoja β anti-paralela en lugar de la hoja paralela nativa; M = 4 (Fig. 15c) revela la ausencia de la hélice 22-35, que es contexto-dependiente, pero posee las restantes características nativas; y finalmente, M = 5 (Fig. 15d) revela un grupo de interacciones terciarias no nativas que protegen fuertemente todas los motivos secundarios. En todas las trayectorias, las estructuras no nativas

son eventualmente desmanteladas en el transcurso de la búsqueda del plegamiento nativo, pero sin comprometer la estabilidad del resto de del núcleo que permanece protegido simultáneamente por otras interacciones.

Una comparación directa del ensamble TS (Fig. 13) para 1gb4 y 1ubi sugiere que, a pesar de que ambas proteínas pertenecen a la misma clase topológica, su diversidad en caminos de plegamiento y distribución en el ensamble TS son dramáticamente distintos. En primer lugar, debería ser señalado que, mientras que la hélice nativa en 1gb4 es muy favorecida por propensiones locales, en 1ubi es extremadamente dependiente del contexto, incluso hasta el punto de requerir la eventual distorsión de estructuras vecinas para prevenir su desmantelamiento (Figs. 12b, 15). La hélice nativa de 1ubi es ausente en el núcleo con frecuencia, en caso de que esté presente en el mismo no puede ser sustentada sin protección terciaria. Todos estos hechos sugieren la existencia de un paso altamente concertado para producir el núcleo, un evento altamente cooperativo. De todas maneras, si el núcleo fuera único, el plegamiento exitoso de 1ubi dependería en un evento muy fortuito, y por lo tanto, no sería reproducible y robusto, como es de esperar de una de las proteínas mas conservadas de la naturaleza, tal como lo es la ubiquitina. Esto implica que esta molécula, producto de la selección natural, debe poseer a su disposición un abanico de posibles rutas de plegamiento alternativas con una diversidad de caminos mucho mayor que la menos conflictiva 1gb4. Justamente, esto es lo que nuestros resultados revelan (Fig. 13): la plasticidad estructural de 1ubi es requerida para garantizar el éxito del proceso de plegamiento dado el conflicto entre propensiones locales (que no favorecen la hélice) y el contexto de largo rango necesario para proteger el núcleo del ataque del agua (el cual favorece de hecho la formación de la hélice, como lo prueba el éxito de las rutas de plegamiento).

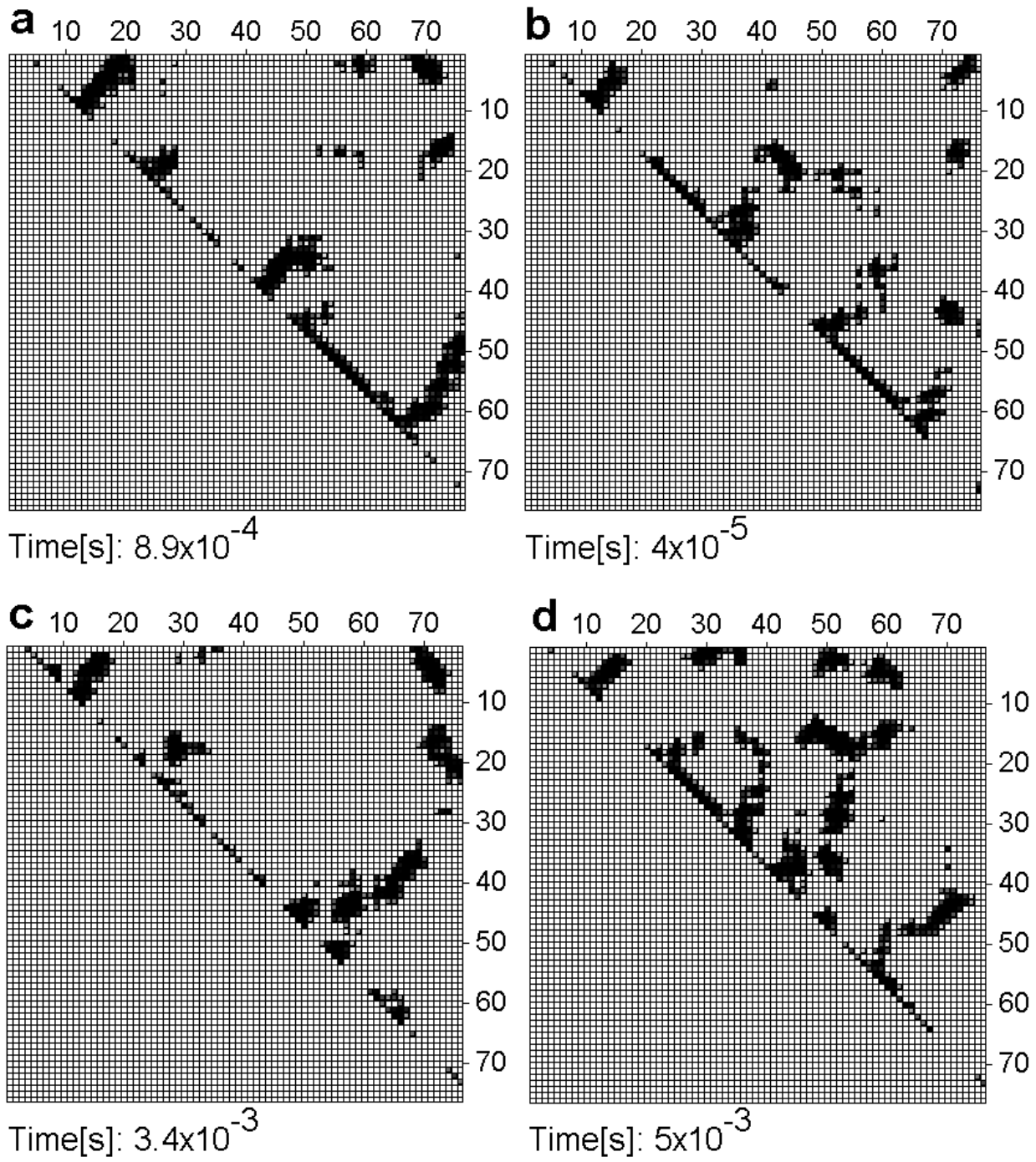


Figura 15: DM's representativas más frecuentes para las componentes M = 2 (a), M = 3 (b), M = 4 (c) y M = 5 (d) del ensamble TS de 1ubi.

5.3 Referencias

[1] D/H amide isotope effects reveal when hydrogen bonds form during protein folding. B. Krantz, L. Moran, A. Kentsis, T. R. Sosnick. *Nature Struct. Biol.* **7**, 62-71 (2000).

[2] The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the protein folding nucleus. Z. Guo, D. Thirumalai. *Fold. Des.* **2**, 377-391 (1997).

[3] Specific nucleus as the transition state for protein folding: evidence from the lattice model. V. I. Abkevich, A. M. Gutin, E. I. Shakhnovich. *Biochemistry* **33**, 10026-10036 (1994).

[4] Finding the collapse-inducing nucleus in a folding protein. A. Fernández, G. Appignanesi, A. Colubri. *J. Chem. Phys* **114**, 8678-8684 (2001).

[5] Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. S. E. Jackson, A. R. Fersht. *Biochemistry* **30**, 10428-10435 (1991).

[6] Molecular collapse: the rate-limiting step in two-state cytochrome c folding. T. R. Sosnick, L. Mayne, S. W. Englander. *Proteins* **24**, 413-426 (1996).

[7] Distinguishing between two-state and three-state models for ubiquitin folding. B. A. Krantz, T. R. Sosnick. *Biochemistry* **39**, 11696-11701 (2000).

[8] Evidence for a 3-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. S. Khorasanizadeh, I. D. Peters, H. Roder. *Nature Str. Biol.* **3**, 193-205 (1996).

[9] Computer-based redesign of a protein folding pathway. S. Nauli, B. Kuhlman, D. Baker. *Nature Str. Biol.* **8**, 602-605 (2001).

[10] Three key residues from a critical contact network in a protein folding transition state. M. Vendruscolo, E. Paci, C. Dobson, M. Karplus. *Nature* **409**, 641-645 (2001).

Capítulo 6

Conclusiones

El objeto de esta tesis es presentar y justificar matemáticamente y físicamente un tratamiento teórico-computacional que entendemos constituye un paso necesario en la solución del problema del plegamiento de una proteína.

Ateniéndonos a las restricciones geométricas y estéricas del movimiento de la cadena proteínica, hemos construido un generador "grosero" de la dinámica torsional que modela la evolución temporal de las restricciones locales en lugar de utilizar una representación en término de coordenadas intrínsecas.

Además por primera vez hemos generado caminos de plegamiento que condujeron al estado nativo corroborable de proteínas naturales. Este hecho nos hace pensar que hemos encontrado el compromiso correcto entre el sacrificio de resolución estructural y la necesidad de abarcar computacionalmente los tiempos biológicamente relevantes, vedados hasta el momento a los tratamientos clásicos de este problema.

Este proceso de simplificación constituye un paso al espacio conformacional cociente, donde dos estados torsionales son vistos como equivalentes si caen en la misma "cuenca de Ramachandran", o sea, si ocupan la misma zona definida por las restricciones locales. Esta dinámica cociente o "módulo cuencas de Ramachandran" encuentra no sólo su justificación matemática en la nueva teoría de procesos estocásticos markovianos propuesta en la tesis, sino también su asidero físico, ya que deriva de la geometría inherente del objeto de estudio.

Apéndice

La máquina de plegar proteínas

En este apéndice se indicará brevemente cómo fue implementado el algoritmo de plegamiento en el programa bautizado "Máquina de Plegar" o FM (Folding Machine).

Desde el punto del vista del usuario, el FM brinda una interfaz intuitiva desde la cual se efectuar, entre otras, las siguientes tareas:

(a) Generar simulaciones de proteínas naturales y diseñadas.

(b) Visualizar la trayectoria de plegamiento utilizando distintos formatos representacionales: vista tridimensional del backbone, matrices de distancias, LTM, etc.

(c) Archivar las simulaciones generadas y recuperarlas más adelante.

(d) Calcular distintos parámetros energéticos y estructurales de una proteína: energía, radio de giro, orden de contacto, etcétera.

(e) Monitorear la evolución de la energía y otros parámetros significativos (número de puentes de hidrógeno, porcentajes de estructura secundaria, etc.) durante el transcurso de una simulación.

(f) Comparar estructuras nativas con las estructuras generadas por la simulación.

(g) Editar la secuencia primaria e introducir mutaciones puntuales arbitrarias.

Por otro lado, la interfase del FM no es estática, sino que se puede ampliar utilizando programas externos, desarrollados por otras personas. En las figuras 16-18 se muestran vistas del FM en funcionamiento, mostrando algunas de las características señaladas recién.

Desde el punto de vista de la implementación, el FM esta compuesto por dos grandes bloques:

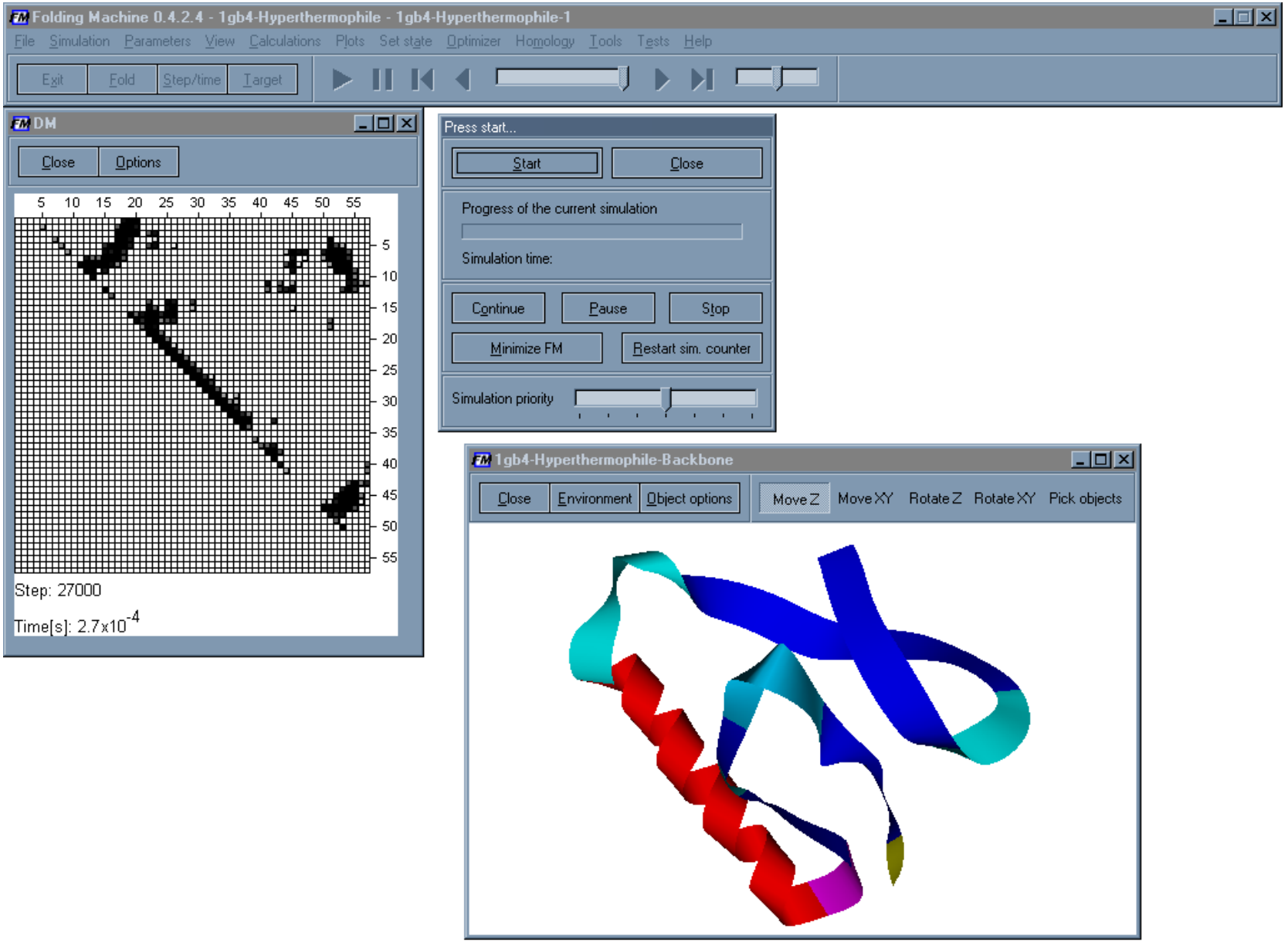
(a) El Folding Machine propiamente dicho, que constituye el GUI (Graphical User Interface) del programa, es decir, la interfase gráfica del mismo.

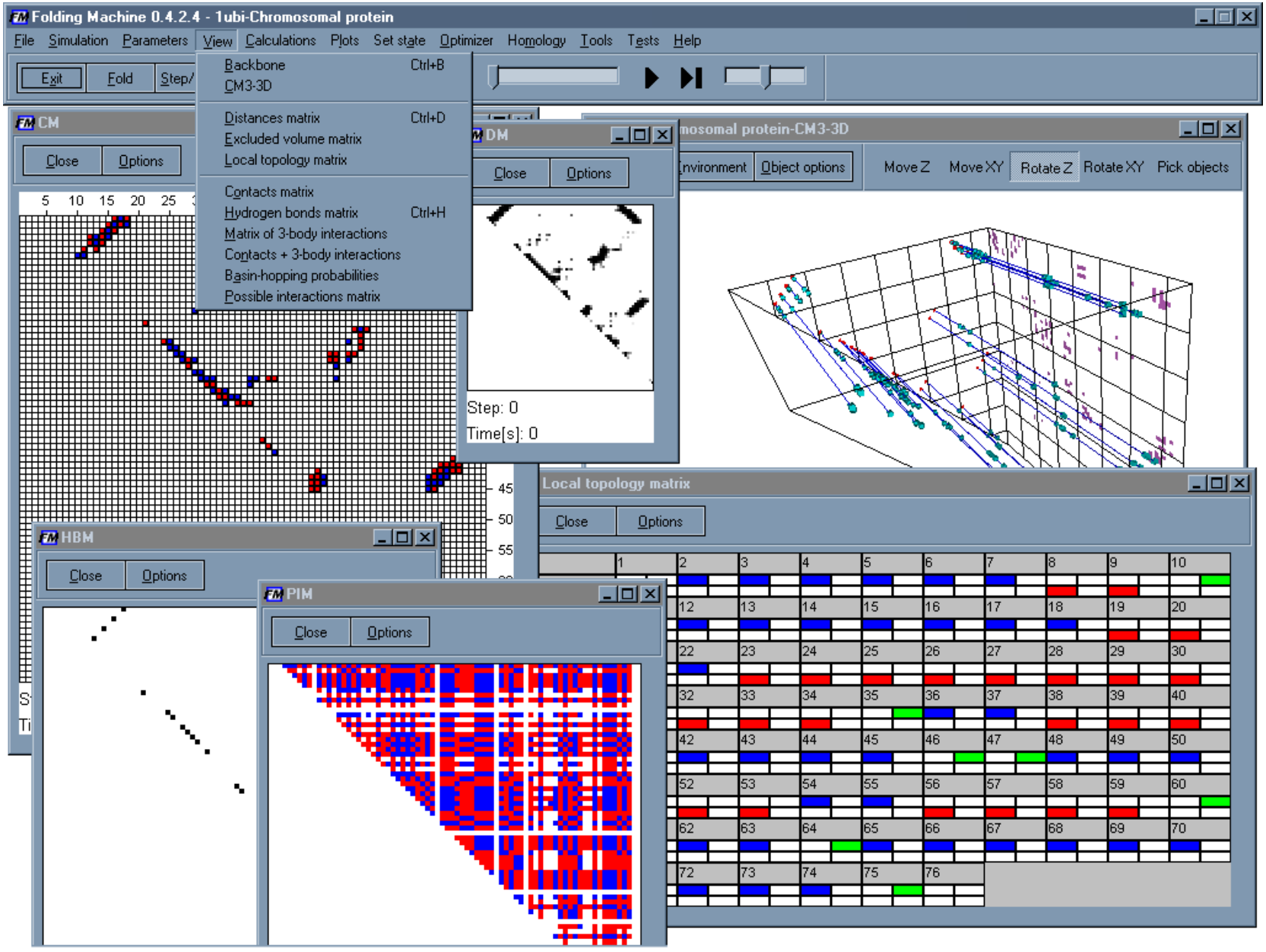
(b) El Folding Engine, que es el núcleo que lleva a cabo todos los cálculos.

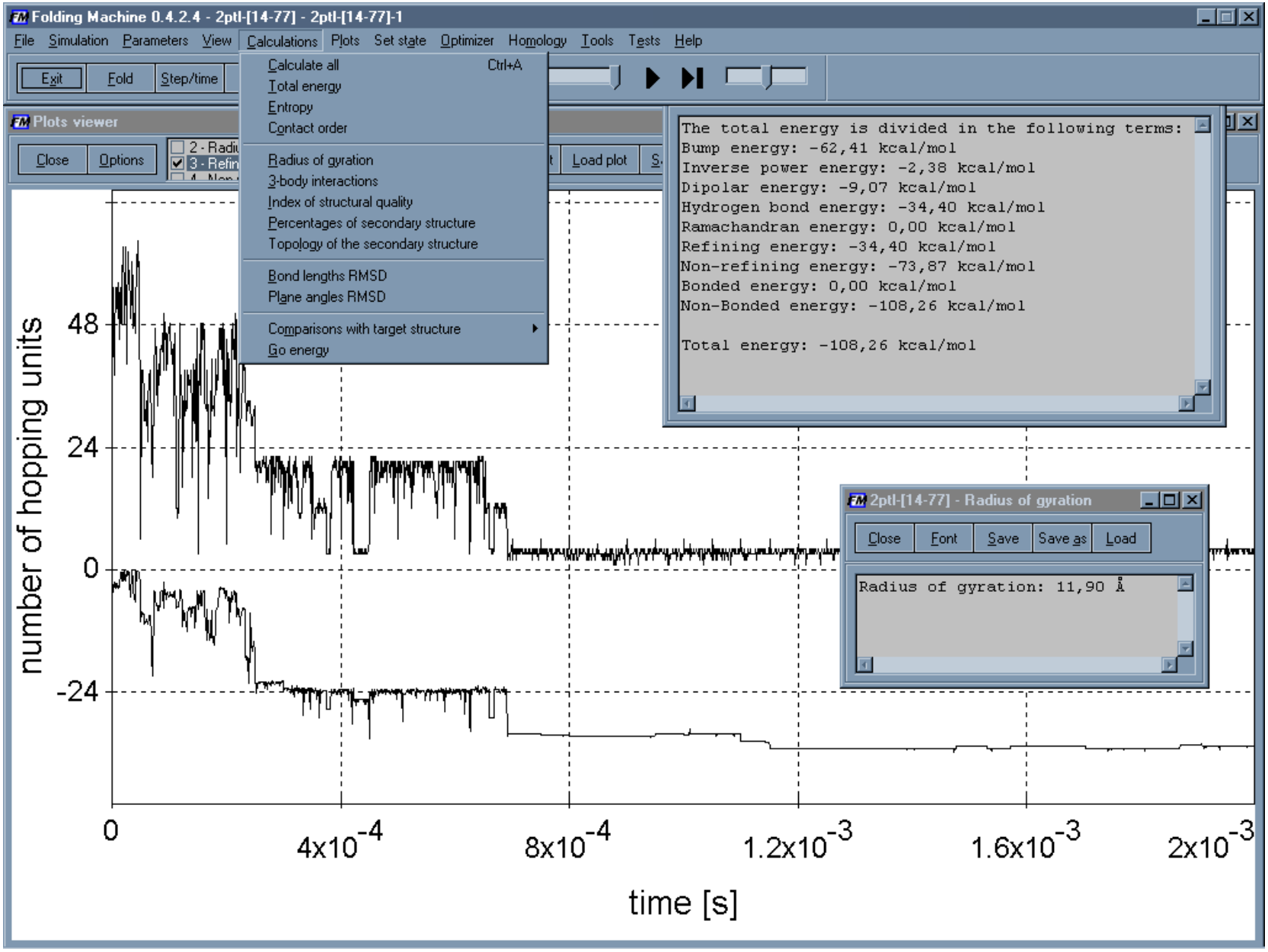
El Folding Engine puede ser pensado como una máquina de estados [1]. Su espacio de estados está formado justamente por las distintas conformaciones que puede adoptar una proteína. El Folding Engine expone en su interfase un conjunto de funciones que definen un auténtico API (Aplication Program Interface) con más de 300 rutinas.

A través de las funciones del API, la conformación interna del Folding Engine puede ser modificada a efectos de generar una nueva conformación. Por ejemplo, una simulación es efectuada llamando reiteradamente a la función que ejecuta un ciclo elemental de simulación a partir de la conformación actual. Gran parte de las operaciones que son realizadas durante el ciclo elemental de simulación pueden ser ejecutadas independientemente, tales como, por ejemplo la optimización de la geometría o el cálculo de las probabilidades de transición entre cuencas de Ramachandran.

Siguientes tres páginas:
Figura 16: Imagen del FM mostrando la interfase principal, la DM , el backbone tridimensional y la ventana de simulación.
Figura 17: Imagen del FM mostrando algunas de las ventanas de visualización de información estructural (Matrix de contactos, LTM, etcétera).
Figura 18: Imagen del FM mostrando la evolución temporal de algunas cantidades (energía total, número de puentes de hidrógeno, etc.) generadas durante una simulación. También se aprecian los resultados de algunos cálculos efectuados sobre la estructura actual.


Folding Machine 0.4.2.4 - 1gb4-Hyperthermophile - 1gb4-Hyperthermophile-1
File Simulation Parameters View Calculations Plots Set state Optimizer Homology Tools Tests Help
Exit Fold Step/time Target
DM
Close Options
5 10 15 20 25 30 35 40 45 50 55
Step: 27000
Time[s]: 2.7x10-4
Press start...
Start Close
Progress of the current simulation
Simulation time:
Continue Pause Stop
Minimize FM Restart sim. counter
Simulation priority
1gb4-Hyperthermophile-Backbone
Close Environment Object options
Move Z Move XY Rotate Z Rotate XY Pick objects


Folding Machine 0.4.2.4 - 1ubi-Chromosomal protein
File Simulation Parameters View Calculations Plots Set state Optimizer Homology Tools Tests Help
Exit Fold Step/
Backbone Ctrl+B
CM3-3D
Distances matrix Ctrl+D
Excluded volume matrix
Local topology matrix
Contacts matrix
Hydrogen bonds matrix Ctrl+H
Matrix of 3-body interactions
Contacts + 3-body interactions
Basin-hopping probabilities
Possible interactions matrix
CM
Close Options
DM
Close Options
Step: 0
Time[s]: 0
nosomal protein-CM3-3D
Environment Object options
Move Z Move XY Rotate Z Rotate XY Pick objects
Local topology matrix
Close Options
HBM
Close Options
PIM
Close Options


Folding Machine 0.4.2.4 - 2ptl-[14-77] - 2ptl-[14-77]-1
File Simulation Parameters View Calculations Plots Set state Optimizer Homology Tools Tests Help
Exit Fold Step/time
Calculate all Ctrl+A
Total energy
Entropy
Contact order
Radius of gyration
3-body interactions
Index of structural quality
Percentages of secondary structure
Topology of the secondary structure
Bond lengths RMSD
Plane angles RMSD
Comparisons with target structure
Go energy
Plots viewer
Close Options
Load plot
The total energy is divided in the following terms:
Bump energy: -62,41 kcal/mol
Inverse power energy: -2,38 kcal/mol
Dipolar energy: -9,07 kcal/mol
Hydrogen bond energy: -34,40 kcal/mol
Ramachandran energy: 0,00 kcal/mol
Refining energy: -34,40 kcal/mol
Non-refining energy: -73,87 kcal/mol
Bonded energy: 0,00 kcal/mol
Non-Bonded energy: -108,26 kcal/mol
Total energy: -108,26 kcal/mol
2ptl-[14-77] - Radius of gyration
Close Font Save Save as Load
Radius of gyration: 11,90 Å
number of hopping units
48
24
0
-24
0
4x10-4
8x10-4
1.2x10-3
1.6x10-3
2x10-3
time [s]

Asimismo, el Folding Engine cuenta con funciones que permiten la entrada y salida de datos (funciones de Entrada/Salida). De esta manera, se pueden cargar estructuras almacenadas en el disco, guardar simulaciones, acceder a las coordenadas torsionales de la conformación actual, etc.

De esta manera, el funcionamiento del FM puede ser descripto de la siguiente manera: cada vez que el usuario selecciona un comando en la interfaz gráfica, el Folding Machine genera el llamado a la función correspondiente del API del Folding Engine. Este llamado modifica consecuentemente el estado interno del Folding Engine y genera eventualmente alguna información de salida, que es tomada por el Folding Machine, procesada y mostrada al usuario.

Podemos representar al FM y las distintas partes que lo forman con el siguiente esquema, donde ↓↑ representa la entrada y salida de información:

Folding Machine GUI (comandos y opciones de visualización)
↓↑
Folding Engine API (funciones de cálculo, E/S, etc.)
↓↑
Funciones numéricas internas
↓↑
Variables de estado almacenadas en la memoria de la computadora
(estado torsional de la proteína, parámetros del sistema, etc.)

Referencias

[1] *Theorical Foundations of Computer Science*. D. Mandrioli, C. Ghezzi. John Wiley and Sons (1987).