UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES “ZARAGOZA”

"Desarrollo de un potencial estadístico basado en la geometría del puente de hidrógeno para la identificación de la estructura nativa de proteínas"

TESIS

QUE PARA OBTENER EL TÍTULO DE:

QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA:

NORBERTO SÁNCHEZ CRUZ

DIRECTOR DE TESIS: DR. RAMÓN GARDUÑO JUÁREZ
ASESOR DE TESIS: DR. JOSÉ IGNACIO REGLA CONTRERAS

FES ZARAGOZA

MÉXICO, D.F. **2014**

JURADO ASIGNADO

PRESIDENTE:	**DRA. MARÍA ISABEL SOTO CRUZ**
VOCAL:	**DR. RAMÓN GARDUÑO JUÁREZ**
SECRETARIO:	**DR. JOSÉ IGNACIO REGLA CONTRERAS**
1er. SUPLENTE:	**DRA. MARTHA LEGORRETA HERRERA**
2do. SUPLENTE:	**DRA. MIRNA RUÍZ RAMOS**

LUGAR DONDE SE DESARROLLÓ EL PROYECTO:

INSTITUTO DE CIENCIAS FÍSICAS DE LA UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

AGRADECIMIENTOS

Al Dr. Ramón Garduño, por permitirme formar parte de su equipo de trabajo, por la confianza y el apoyo brindados durante el desarrollo de este proyecto.

Al honorable jurado, por sus valiosas enseñanzas y aportaciones para enriquecer este trabajo.

Al cDr. Gilberto Sánchez y al M. en C. José Luis Velasco, por sus asesorías y enseñanzas, me ayudaron mucho.

A la Facultad de Estudios Superiores “Zaragoza”, por permitirme formar parte de sus filas y crecer en sus aulas y laboratorios. Es un honor ser egresado esta casa de estudios.

A la Universidad Nacional Autónoma de México, que durante 8 años se ha encargado en gran medida de mi formación, y que con sus distintos programas de apoyo me ha facilitado enormemente el camino.

A mis padres, David y Magdalena, por cuidarme, apoyarme y darme las armas para forjar mi propio camino, creer en mí, enseñarme que puedo llegar tan lejos como me lo proponga y darme todo lo que ha estado en sus manos para que pueda lograrlo.

A mi abuelita Martina y mi tía Leonila, por el esfuerzo que han hecho por sacar a su familia adelante y estar siempre al pendiente de mí.

A mis hermanos: Edith, David, Jesús y Román, sus parejas y cada uno de mis sobrinos, por el apoyo incondicional, por estar ahí siempre que los he necesitado, por sus consejos y enseñanzas a lo largo de toda mi vida, por darme una razón más para esforzarme por ser mejor y por llenar mi vida de felicidad. Simplemente no podría tener una familia mejor y sé que sin ustedes no sería la persona que soy.

A mis amigos: Eduardo, por acompañarme desde el primer día que pisé esta universidad hasta... seguimos contando. A Jessica, por todas las experiencias compartidas, por hacerme crecer profesional y personalmente. A Julio y Sandra, por su constante apoyo y compañía durante toda la carrera. A Isaac, Fabián y Luis, por volverse como mis hermanos durante nuestra estancia en Cuernavaca. A Beto, Karen, Diego, Fili, Leonel, Jaime, Chucho... seguro me faltan varios, por todos esos buenos momentos que pasamos juntos durante todos estos años. A los amigos que conocí en Cuernavaca, especialmente Adriana y David, por hacerme sentir como en casa. Cada uno de ustedes ocupa un lugar muy especial en mi vida.

A los compañeros y profesores, de los que he aprendido tanto a lo largo del camino, especialmente a Jorge Rivas y Patricia Vidal, dos grandes personas que marcaron etapas muy importantes en mi desarrollo tanto profesional como personal.

DEDICATORIA

Con todo mi cariño para las personas que han hecho todo para permitirme alcanzar mis sueños, por darme la mano y no dejarme caer en los momentos más difíciles.

Mamá, papá, hermanos: los amo.

ÍNDICE GENERAL

ABREVIATURAS

A	Aceptor
Å	Angstroms
AMBER	Assisted Model Building with Energy Refinement
D	Donador
g	Gramos
H	Hidrógeno
L	Litros
N	Nitrógeno
O	Oxígeno
PBC	Potencial Basado en el Conocimiento
PDB	Protein Data Bank
u.a.	Unidades arbitrarias
UniProt	Universal Protein Resource
δ	Distancia Hidrógeno-Aceptor
θ	Ángulo Donador-Hidrógeno-Aceptor

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Objetivo: Desarrollar un potencial estadístico para la identificación de la estructura nativa de las proteínas, basado en la geometría del puente de hidrógeno en estas.

Material y métodos: Utilizando una muestra de 2078 estructuras de proteínas en su conformación nativa obtenidas del Protein Data Bank, se construyeron potenciales estadísticos basados en dos parámetros geométricos de los puentes de hidrógeno intramoleculares formados por la cadena principal de estas, la distancia hidrógeno-aceptor y el ángulo donador-hidrógeno-aceptor. Fueron construidos dos potenciales distintos, variando el intervalo de medición del ángulo donador-hidrógeno-aceptor para evaluar la influencia de los puentes de hidrógeno con dicho ángulo entre 90 y 120 grados en su capacidad de discriminación de la estructura nativa en un conjunto de confórmeros. Para evaluar dicha capacidad, se les probó en grupos de confórmeros de 76 proteínas distintas.

Resultados: Se encontró, con un 95% de confianza, que la medición de los puentes de hidrógeno con un ángulo donador-hidrógeno-aceptor juega un papel importante en la capacidad del potencial para discriminar la estructura nativa de una proteína en un conjunto de confórmeros, ya que incrementa su poder predictivo de 65% a 70%.

Conclusión: La capacidad de un potencial estadístico basado en dos descriptores, uno de distancia y otro de direccionalidad, para los puentes de hidrógeno formados por la cadena principal de proteínas con estructura terciaria conocida, se encuentra alrededor del 70% para predecir la estructura nativa de una proteína en un conjunto de señuelos, siendo los puentes de hidrógeno con un ángulo entre 90° y 120° un factor importante para la realización de dicha discriminación.

I. Introducción

Conocer la estructura terciaria de una proteína es importante para entender diversos aspectos de su función y poder emplear dicho conocimiento en el diseño de nuevos fármacos [1,2]. Sin embargo, en la base de datos del Universal Protein Resource (UniProt) se tienen almacenadas más de 48 millones de secuencias de aminoácidos de proteínas [3], de las cuales solo alrededor de 95 mil poseen una estructura tridimensional conocida y almacenada en la base de datos del Protein Data Bank (PDB)[4]. Es por ello que uno de los problemas actuales más retadores en el campo de la biología molecular computacional es el de predecir la estructura nativa de una proteína con base a su secuencia de aminoácidos**.** Los métodos empleados en la resolución de este problema se pueden clasificar en tres categorías [5]: modelado comparativo, reconocimiento del plegado y métodos *ab initio*. Los primeros dos grupos predicen estructuras de proteínas basados en estructuras de proteínas ya resueltas [6-12], mientras que los que comprenden el tercer grupo se basan en la hipótesis termodinámica, la cual establece que la estructura nativa de una proteína será aquella que posea la menor energía libre bajo condiciones fisiológicas [13], lo cual implica la necesidad de desarrollar funciones de energía que permitan la identificación de dicha estructura. Dentro de estas funciones se encuentran los potenciales basados en el conocimiento [14-17] (PBC) o potenciales estadísticos**,** los cuales son funciones energéticas derivadas de bases de datos de estructuras de proteínas cuya estructura tridimensional es conocida [18]. Los PBC capturan, empíricamente, los aspectos más relevantes de la fisicoquímica de la estructura y función de las proteínas. Son derivados al medir la probabilidad de una observable en un conjunto de estructuras experimentales en relación a un estado de referencia y la conversión de la probabilidad a una función de energía se realiza normalmente empleando la ley de

Boltzmann. Los PBC han tenido éxito en la predicción de la estructura de proteínas, en la predicción de las interacciones proteína-proteína, en la predicción de las interacciones ligando-proteína y en el diseño de proteínas. Sin embargo, la mayoría de los PBC explotan parámetros observables como la distancia entre pares o cuartetos de los carbonos alfa, o de los centroides de las cadenas laterales, que resultan útiles cuando no se requiere una predicción en detalle atómico.

Por otro lado, existen evidencias experimentales y computacionales que confirman el papel esencial de preferencias locales en la configuración de las estructuras proteicas. Entre ellas se encuentra el puente de hidrógeno intramolecular que es uno de los tipos de interacción molecular no covalente más importante en biología, y al cual se le atribuye la capacidad de conferir la orientación y especificidad de las interacciones intramoleculares [19].

Nuestra hipótesis de trabajo es la de explorar la posibilidad que un PBC basado en la direccionalidad de los puentes de hidrógeno intramoleculares de proteínas plegadas en su estructura nativa es suficiente para distinguir a esta en un conjunto de diferentes estructuras con la misma secuencia primaria, conocidas como señuelos. Por tal motivo, este trabajo se centrará en el desarrollo de un potencial estadístico basado en el análisis geométrico de los puentes de hidrógeno formados por la cadena principal de proteínas con estructura terciaria conocida, que posteriormente pudiera ser empleado como función evaluadora para el desarrollo de algoritmos de predicción de estructura terciaria de proteínas que a su vez pueden ser usados como herramienta para la síntesis racional de fármacos.

II. Marco teórico

1. Proteínas y sus niveles de organización estructural

Las proteínas son el mayor componente de los organismos vivos, ya que ocupan alrededor del 50% del peso seco de una célula, más que cualquier otra biomolécula. Son también las macromoléculas encargadas de permitir que se lleven a cabo prácticamente todas las reacciones que ocurren en un sistema biológico. Las proteínas son polímeros de aminoácidos, sin embargo, de entre todas las posibilidades de aminoácidos que pueden existir, solo veinte de ellos se encuentran en la mayoría de las proteínas, estos son: alanina, valina, leucina, isoleucina, glicina, prolina, cisteína, metionina, histidina, fenilalanina, tirosina, triptófano, asparagina, gltamina, serina, treonina, lisina, arginina, ácido aspártico y ácido glutámico [20].

Para su estudio, las proteínas se pueden examinar en cuatro distintos niveles de estructura: primaria, secundaria, terciaria y cuaternaria [20-23].

1.1. Estructura primaria

La estructura primaria es la secuencia exacta de aminoácidos que forman su cadena. Esta secuencia es muy importante, ya que de ella depende el plegado final, y por tanto, la función de cada proteína. Cada uno de estos aminoácidos posee una estructura que incluye un carbono alfa rodeado de cuatro sustituyentes: un hidrógeno, un grupo amino, un grupo carboxilo y una cadena lateral (Figura 1).

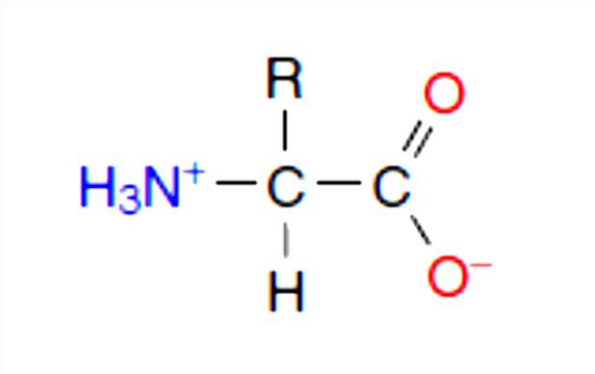


Figura 1. Estructura general de un aminoácido. En azul se muestra el grupo amino y en rojo el grupo carboxilo [20].

A su vez, la estructura primaria puede considerarse como dos cadenas distintas, la primera formada por el carbono alfa, el grupo amino y el grupo carboxilo, la cual es idéntica para todos los aminoácidos y se conoce como cadena principal o esqueleto; y la segunda formada por la cadena lateral y es específica para cada aminoácido, por tanto la encargada de diferenciarlos. De acuerdo a sus características, las cadenas laterales se pueden clasificar como: alifáticas, no polares, aromáticas, polares y cargadas. La secuencia de aminoácidos en una proteína se lee desde su extremo amino hacia su extremo carboxilo, a cada aminoácido se le ha asignado un código de tres letras que lo identifica, pero con el fin de ahorrar espacio en la representación de secuencias de proteínas muy largas, se usa un código de una letra para cada aminoácido (Tabla 1).

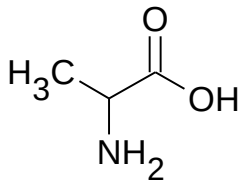
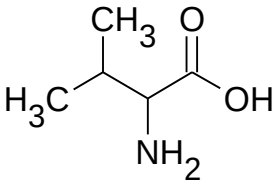
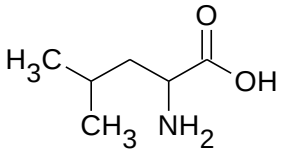
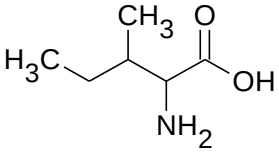
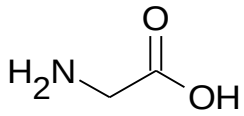
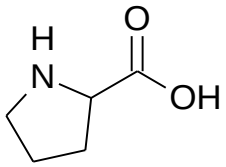
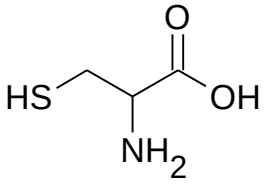
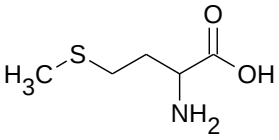
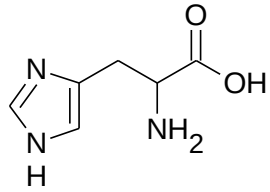
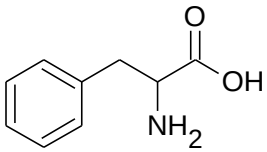
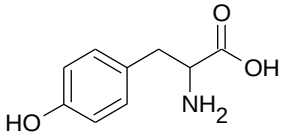
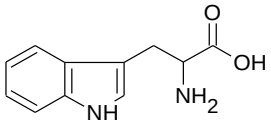
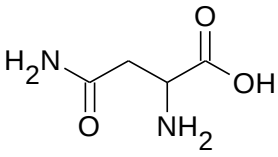
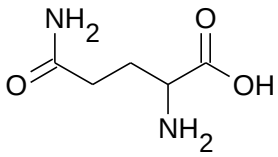
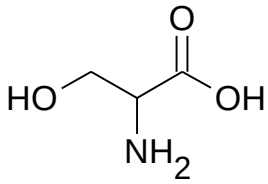
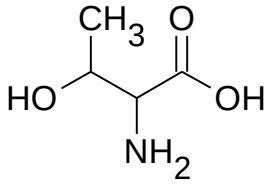
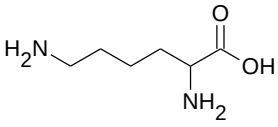
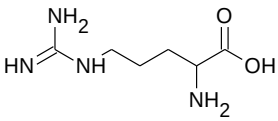
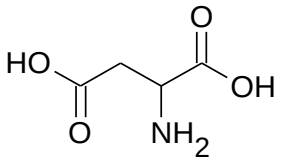
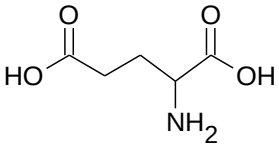
Aminoácidos con cadena lateral alifática			
Alanina (Ala, A)	Valina (Val, V)	Leucina (Leu, L)	Isoleucina (Ile, I)
			
Aminoácidos con cadena lateral no polar			
Glicína (Gly, G)	Prolina (Pro, P)	Cisteína (Cys, C)	Metionina (Met, M)
			
Aminoácidos con cadena lateral aromática			
Histidina (His, H)	Fenilalanina (Phe, F)	Tirosina (Tyr, Y)	Triptófano (Trp, W)
			
Aminoácidos con cadena lateral polar			
Asparagina (Asp, N)	Glutamina (Gln, Q)	Serina (Ser, S)	Treonina (Thr, T)
			
Aminoácidos con cadena lateral cargada			
Lisina (Lys, K)	Arginina (Arg, R)	Ácido aspártico (Asp, D)	Ácido glutámico (Glu, E)
			

Tabla 1. Estructuras de los aminoácidos constituyentes de las proteínas. Organizados según las propiedades de su cadena lateral, en paréntesis los códigos de tres y una letras de cada uno [23]

A excepción del aminoácido más pequeño, la glicina, que en lugar de una cadena lateral lleva otro átomo de hidrógeno, los otros 19 aminoácidos poseen cuatro sustituyentes diferentes, lo que significa que su carbono alfa es un centro quiral, por lo que pueden presentarse dos enantiómeros: D y L; sin embargo los procesos de síntesis biológica de proteínas generan sólo proteínas constituidas exclusivamente por L-aminoácidos (Figura 2).

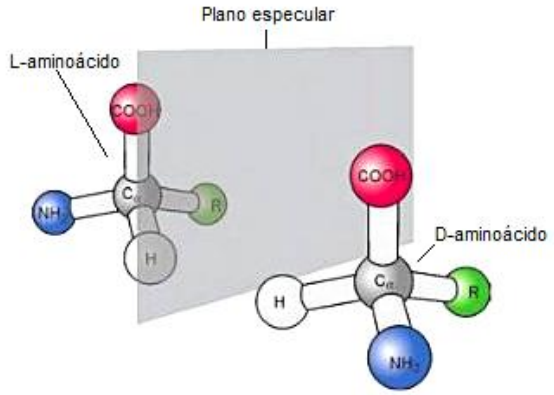


Figura 2. Formas enantioméricas de los α-aminoácidos. El centro quiral es en cada caso el carbono α [22].

Los aminoácidos que conforman una proteína se mantienen unidos por medio de la formación de enlaces peptídicos, en ellos el grupo amino de un aminoácido reacciona con el grupo carboxilo de otro. Esta reacción es descrita como una condensación, resultando en la eliminación de una molécula de agua (Figura 3).

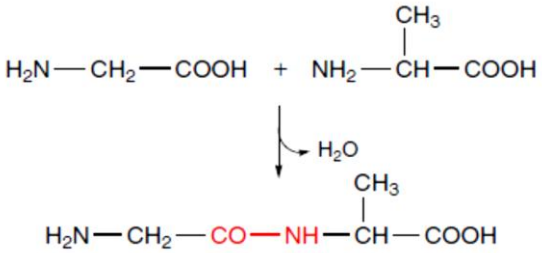


Figura 3. Formación del enlace peptídico entre dos aminoácidos (glicína y alanina) [20].

1.2. Estructura secundaria

La estructura primaria conduce a la estructura secundaria, la cual se refiere a la conformación local o relación espacial que existe entre aminoácidos que se encuentran cercanos en la estructura primaria. Las unidades básicas de la estructura secundaria son: las hélices alfa, las hebras beta y los giros (o vueltas de horquilla); el resto de las unidades de estructura secundaria conocida son variaciones de estos tres tipos. Todas las unidades de estructura secundaria mencionada se generan por la formación de puentes de hidrógeno entre los distintos grupos carboxilo y amino de residuos diferentes [20-22].

1.2.1. Hélice alfa

La hélice alfa es una estructura de forma helicoidal que se genera por la torsión uniforme de la cadena polipeptídica. Esta hélice es dextrógira: gira en el sentido de las manecillas del reloj si se mira a lo largo de su eje en dirección carboxilo terminal-amino terminal. Los puentes de hidrógeno entre pares de residuos están distanciados entre sí por tres residuos y su repetición es periódica a lo largo de toda la hélice, las cadenas laterales se orientan hacia el exterior de la hélice (Figura 4).

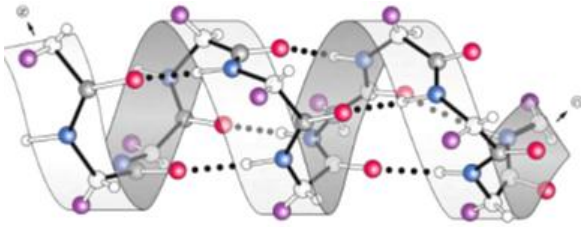


Figura 4. Representación de una hélice alfa. Se muestran todos los átomos de la cadena principal (carbono alfa: blanco, oxígeno: rojos, nitrógeno: azul), las cadenas laterales son representadas por esferas moradas [22].

1.2.2. Hoja beta plegada

Las hojas beta plegadas son unidades de estructura secundaria formadas por varias hebras beta. En estas estructuras, a diferencia de la hélice alfa, no interaccionan segmentos continuos de una sola cadena polipeptídica, sino diferentes combinaciones de secciones no necesariamente inmediatas una a la otra y que pueden o no pertenecer a la misma cadena polipeptídica. Las hebras beta están dispuestas una junto a la otra de tal manera que se pueden formar puentes de hidrógeno entre los grupos carboxilo de una hebra y los grupos amino de otra [20-22].

Las dos hebras que interaccionan pueden ser paralelas (en el mismo sentido) o antiparalelas (en sentidos opuestos). Ambas cadenas se pliegan en forma de acordeón y las cadenas laterales se posicionan de forma alterna por encima y por debajo de la hoja beta (Figura 5).

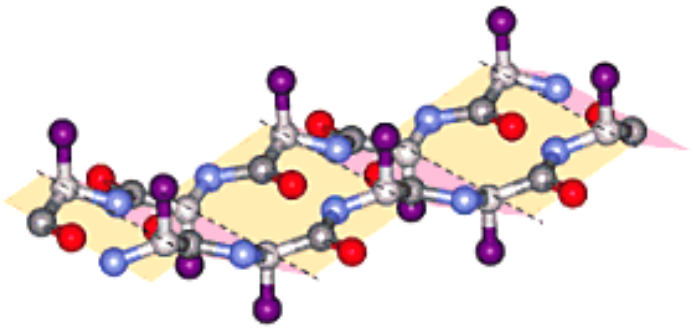


Figura 5. Representación de una hoja beta plegada de dos hebras. Código de colores como en la figura 4 [22].

1.2.3. Giros (vueltas de horquilla)

Las vueltas de horquilla son el tercer elemento clásico de estructura secundaria, estos se encargan de conectar entre si los otros elementos de estructura secundaria. Los giros más comunes son los beta, que son aquellos que unen dos hebras en una hoja beta plegada. Existen diversos tipos de giros beta, los más frecuentes son los “tipo I” y “tipo II”, cada uno de ellos contiene 4 residuos que se estabilizan por la formación de un puente de hidrógeno entre los grupos carboxilo y amino de los residuos 1 y 4 respectivamente, mientras que los residuos 2 y 3 varían en su conformación, lo que da origen a los distintos tipos de giro (Figura 6) [20-22].

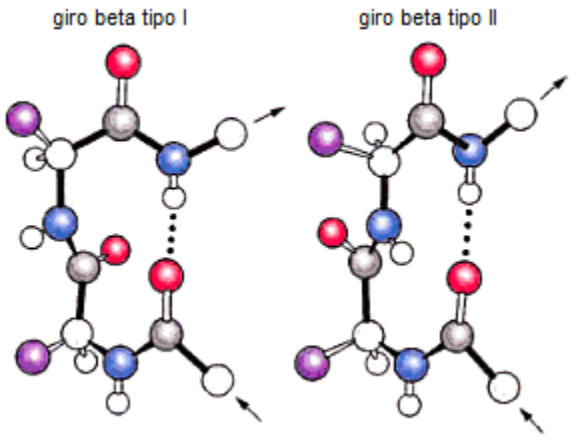


Figura 6. Representación de 2 tipos de giros beta. Código de colores como en la figura 4 [22].

1.3. Estructura terciaria

La estructura terciaria representa la secuencia completa de aminoácidos totalmente plegada, se define como el arreglo espacial de los aminoácidos que se encuentran ampliamente separados en la estructura primaria, o de manera más concisa, como la topología total formada por la cadena polipeptídica.

La formación de estructura terciaria estable depende de las interacciones que se presentan en la molécula, las cuales difieren tanto en fuerza como en frecuencia de ocurrencia. Ejemplos de estas interacciones son los puentes disulfuro, el efecto hidrofóbico, interacciones coulombicas, puentes de hidrógeno e interacciones de Van der Waals [20-23].

1.4. Estructura cuaternaria

Muchas proteínas contienen más de una cadena polipeptídica. La interacción entre estas cadenas es lo que da origen a la estructura cuaternaria. Las interacciones que propician la formación de estructura cuaternaria en una proteína son exactamente las mismas que originan la estructura terciaria, con la excepción de que estas ocurren entre dos o más cadenas polipeptídicas.

2. Interacciones que estabilizan la estructura de las proteínas

Como se mencionó anteriormente, existen diversos tipos de interacción que se encargan de estabilizar las estructuras terciaria y cuaternaria de las proteínas, a continuación se describen brevemente cada una de estas interacciones [19,20].

2.1. Puentes disulfuro

Los puentes disulfuro dictan el plegado de las proteínas por formación de fuertes enlaces covalentes entre los átomos de azufre de las cadenas laterales de cisteínas que se encuentren a más de cinco aminoácidos de separación. Un puente de este estilo no se puede formar entre residuos consecutivos, comúnmente los residuos de cisteínas se encuentran separados por al menos 5 residuos. Estos enlaces pueden romperse a alta temperatura, pH ácido o en presencia de agentes reductores.

2.2. El efecto hidrofóbico

La escasa solubilidad de las moléculas no polares en agua provoca que, en un ambiente acuoso, las interacciones de este tipo de moléculas entre si se magnifiquen. Como la mayoría de las proteínas se encuentran en un ambiente de este estilo, ocurre la formación de aglomerados de aminoácidos con cadena lateral no polar, entre los cuales el agua queda excluida, dejando en claro que este efecto contribuye significativamente al total de las interacciones intramoleculares presentes en una proteína [20,23].

2.3. Interacciones coulombicas

Este tipo de interacciones se dan entre aminoácidos con cadenas laterales cargadas, así como los grupos amino (NH3 +) y carboxilo (COO -) terminales de las cadenas polipeptídicas, y se describen en base a la ley de Coulomb. Como resultado los aminoácidos con cadenas laterales cargadas se encuentran normalmente en la superficie de las proteínas estabilizando la estructura, y su interacción con aminoácidos vecinos se debilita por la presencia de las moléculas de agua con un efecto de apantallamiento [20,23].

2.4. Interacciones de Van der Waals

Existen fuerzas de Van der Waals tanto atractivas como repulsivas que controlan las interacciones entre átomos sin carga y no enlazados entre sí. Estas fuerzas provienen de la inducción de dipolos en la molécula debido a la fluctuación de las densidades de carga entre los átomos. Las interacciones englobadas en este efecto y ordenadas de mayor a menor fuerza, son: la interacción entre dipolos permanentes, la interacción entre dipolos temporales y las fuerzas de dispersión de London [20,23].

2.5. Puentes de hidrógeno

Los puentes de hidrógeno contribuyen significativamente a la estabilidad de las hélices alfa y a la interacción de las hebras beta para la formación de hojas beta. Como resultado de ambos fenómenos, el puente de hidrógeno contribuye significativamente a la estabilidad total de la estructura terciaria de una proteína. La mayor parte de estas interacciones está dada por puentes de hidrógeno formados entre los grupos NH y CO de distintos enlaces peptídicos en la cadena principal, aunque también pueden existir puentes de hidrógeno entre las cadenas laterales de distintos aminoácidos, o entre la cadena principal de la proteína y las cadenas laterales de algunos residuos.

En todos los casos, un puente de hidrógeno involucra un átomo donador y un átomo aceptor, específicamente en el caso de los puentes de hidrógeno entre la cadena principal de una proteína, el átomo donador es el nitrógeno del grupo NH, mientras que el átomo receptor es el átomo de oxígeno del grupo CO.

Un estudio más detallado de la geometría del puente de hidrógeno puede ser encontrado en el trabajo realizado por Baker and Hubbard [24], en el que describen que la distancia entre el

átomo de hidrógeno y el átomo aceptor de los puentes de hidrógeno en proteínas oscila entre 1.6 y 2.5 Angstroms, mientras que el ángulo formado por el triplete donador-hidrógeno-aceptor raramente se encuentra por debajo de 120°.

3. Plegamiento de proteínas

Una gran cantidad de proteínas poseen un plegamiento espontáneo *in vitro*, confirmando la idea propuesta por Anfinsen [13], que propone que la secuencia lineal de aminoácidos en una cadena polipeptídica contiene toda la información necesaria para dictar la estructura tridimensional de las proteínas. A pesar de que el plegado de proteínas se ha estudiado ampliamente por más de 50 años, la explicación de cómo ocurre este proceso sigue siendo uno de los mayores problemas en biología. Por otra parte, se sabe que en la célula, una gran proporción de proteínas, al ser sintetizadas, requieren de la ayuda de chaperonas moleculares para alcanzar su estructura nativa eficientemente.

3.1. Plegamiento in vitro

Para las proteínas conformadas por más de 100 aminoácidos (aproximadamente el 90% de las proteínas en una célula), se considera casi por regla, que alcanzan su conformación nativa pasando por distintos intermediarios de plegado, debido a la mayor tendencia que poseen para colapsar rápidamente en disoluciones acuosas y formar conformaciones compactas no-nativas. Sin embargo, se ha demostrado que incluso proteínas pequeñas deben pasar a través de intermediarios estructurales en su camino hacia la conformación nativa [25,26]. Dichos intermediarios representan conformaciones no plegadas pero cinéticamente estables que requieren reorganizarse antes de alcanzar la conformación nativa. Desde la perspectiva de la teoría del paisaje energético del plegado, las proteínas tienen un

paisaje en forma de embudo, dirigido hacia la conformación nativa, la cual, de acuerdo con la hipótesis de Anfinsen [14] corresponde a la conformación de menor energía libre [27,28] (Figura 7). La presencia de estos intermediarios cinéticamente estables explica la rugosidad del embudo de plegado.

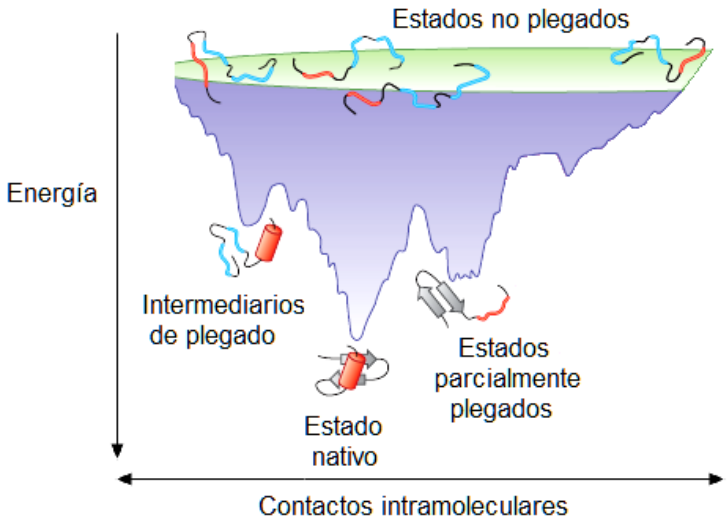


Figura 7. Representación esquemática del embudo de plegado [26].

3.2. Plegamiento in vivo

En la célula, la formación de intermediarios de plegado se dificulta debido a la gran cantidad de moléculas presentes en el medio, cuya concentración es aproximadamente 350 g/L, lo cual promueve las interacciones entre distintas macromoléculas que conllevarían a la formación de agregados entre ellas. Adicionalmente, el proceso de traducción de una proteína incrementa por sí mismo la probabilidad de un mal plegamiento debido a que cadenas polipeptídicas incompletas no pueden formar los intermediarios estables para alcanzar la conformación nativa de la proteína completa, por lo tanto, sería necesario que se

sintetizara toda la proteína, o al menos una subunidad completa, antes de empezar el proceso de plegado.

Es en este punto que intervienen las chaperonas moleculares, que son proteínas que interactúan de manera paralela al proceso de traducción de una proteína con la finalidad de inhibir su plegamiento prematuro. Cabe resaltar que este tipo de proteínas no interactúan aportando información al proceso de plegado, simplemente contribuyen a la optimización de dicho proceso. Por otro lado, algunas proteínas poseen bajas eficiencias intrínsecas de plegado y esencialmente no pueden plegarse en ausencia de chaperonas debido a que no pueden alcanzar la energía suficiente para modificar una conformación cinéticamente estable [27,28].

3.3. Importancia de la predicción del plegado de proteínas

Conocer la estructura terciaria de una proteína es importante para entender diversos aspectos de su función y poder emplear dicho conocimiento en el diseño de nuevos fármacos [1,2]. Sin embargo, en la base de datos de UniProt se tienen almacenadas más de 48 millones de secuencias de aminoácidos de proteínas [3], de las cuales, sólo alrededor de 95 mil poseen una estructura terciaria descrita y almacenada en la base de datos del PDB [4]. Por lo tanto, la predicción de la estructura terciaria de una proteína en base a su secuencia de aminoácidos es uno de los problemas actuales más retadores, y su resolución entra en el campo de la biología molecular computacional [35].

4. Métodos empleados en la predicción del plegado de proteínas

Los métodos empleados en la predicción de la estructura terciaria de proteínas con base a su secuencia de aminoácidos se pueden clasificar en tres categorías [5]: modelado comparativo, reconocimiento del plegado y métodos ab initio; estos se describen a continuación.

4.1. Modelado comparativo

De los datos experimentales disponibles se ha observado que proteínas con secuencias similares de aminoácidos tienden a adoptar estructuras terciarias similares. Es por ello que la manera más sencilla de predecir la estructura terciaria de una proteína es construyendo una estructura en base a proteínas conocidas que compartan gran parte de la secuencia con la proteína de estudio. En muchos de estos casos las proteínas de comparación pertenecen a la misma familia biológica que la proteína de estudio, por lo cual este tipo de métodos es conocido también como modelado por homología [5].

4.2. Reconocimiento del plegado

De todas las proteínas con secuencia conocida, solo del 10 al 20% se pueden modelar por homología, para el resto es necesario recurrir a otros métodos. El reconocimiento del plegado se basa en la suposición de que el número de pliegues existentes en las proteínas se limita a alrededor de 1000 [29], y la meta de estos métodos es reconocer cuál de estos plegados corresponde a la estructura nativa de la proteína en cuestión, basándose en las tendencias de ciertos aminoácidos para formar determinadas estructuras secundarias [5].

4.3. Métodos ab initio

A pesar de los grandes esfuerzos realizados, cuando se emplean los métodos ya mencionados, existe un gran número de secuencias de proteínas cuya estructura terciaria no

se ha podido modelar adecuadamente. Es por ello que el uso de métodos *ab initio* se vuelve indispensable. Estos métodos consisten en explorar el espacio conformacional de una proteína hasta encontrar el estado con la menor energía libre. Todos los métodos ab initio poseen tres componentes esenciales [30]:

- *Una forma de modelar a las proteínas:* la cual puede ir desde la descripción detallada de cada átomo hasta modelos más simplificados donde un aminoácido completo se considera como una sola partícula.

- *Un método de búsqueda en el espacio conformacional:* es un problema combinatorio, se puede emplear cualquier algoritmo útil en la resolución de este tipo de problemas [31]

- *Una función que asigne un valor de energía a cada estado conformacional*: es una función algebraica que sirve como punto de comparación para detectar el estado de menor energía libre [14-17]. Dentro de estas funciones se encuentran los campos de fuerza y los PBC o potenciales estadísticos [18].

4.3.1. Campos de fuerza

Son funciones de energía que describen las interacciones entre todos los componentes de un sistema molecular en términos de mecánica molecular, que contiene parámetros de acuerdo a los tipos de enlace, longitudes de enlace, ángulos de enlace, ángulos diedros, cargas y otras interacciones electrostáticas [14-17]. En general este tipo de funciones son las que mejor describen las propiedades fisicoquímicas de las moléculas, pero también corresponden a los cálculos más tardados.

4.3.2. Potenciales estadísticos

Son funciones energéticas derivadas de bases de datos de estructuras tridimensionales de proteínas conocidas [18]. Existe una gran variedad de funciones de este estilo, en seguida se mencionan brevemente algunos que han contribuido de manera importante al desarrollo de esta área:

- La aproximación cuasi-química de Miyazawa y Jernigan [32] que representa a cada residuo por el centroide de su cadena lateral y asigna una energía para cada contacto entre aminoácidos cercanos que no sean vecinos en secuencia, suponiendo que las proteínas son un sistema canónico que sigue la distribución de Boltzmann y que se encuentran en un estado de cuasi-equilibrio químico. Este tipo de potencial tiene una capacidad predictiva del 70%.

- El potencial de fuerza media de Sippl [33] que genera una función de energía para cada contacto entre pares de aminoácidos en base a las distancias entre todos los átomos presentes en dicho contacto. Este potencial tiene una capacidad predictiva del 80%.

- Los potenciales de cuatro cuerpos, inicialmente calculados por Krishnamoorthy y Tropsha [34], superan a los de dos cuerpos en su poder predictivo, alcanzando un 90%.

- El potencial de cuatro cuerpos basado en Beta-Complex [35], que usa la teoría de Beta-Shape y la cuasi-triangulación, con una capacidad predictiva del 95%.

III. Planteamiento del problema

La estructura terciaria de una proteína es determinante para su función biológica. El entender los factores que intervienen en este proceso es de suma importancia en farmacología. Sin embargo, se conoce la estructura terciaria de un número pequeño de proteínas en comparación con la cantidad de secuencias que se han reportado. Para coadyuvar en esta tarea es necesario el desarrollo de una herramienta que ayude en la predicción de dichas estructuras. Dado que los puentes de hidrógeno son una de las interacciones más importantes en la estabilización de la estructura terciaria de las proteínas, el presente trabajo se basa en el desarrollo de un potencial estadístico basado en la geometría del puente de hidrógeno para la identificación de la estructura nativa de estas, y que de resultar útil, puede ser empleado como función evaluadora para el desarrollo de algoritmos de predicción de estructura terciaria de proteínas que a su vez pueden servir como herramienta para la síntesis racional de fármacos.

¿Un potencial estadístico basado en la geometría del puente de hidrógeno en proteínas será suficiente para reconocer la estructura nativa de estas en un conjunto de diferentes confórmeros con la misma secuencia primaria?

IV. Hipótesis

Un potencial estadístico basado en la geometría del puente de hidrógeno en proteínas es suficiente para reconocer la estructura nativa de estas en un conjunto de diferentes estructuras con la misma secuencia primaria, conocidas como señuelos.

V. Objetivos

1. Objetivo general

- Desarrollar un potencial estadístico para la identificación de la estructura nativa de las proteínas, basado en la geometría del puente de hidrógeno en estas.

2. Objetivos particulares

- Llevar a cabo la revisión bibliográfica pertinente para el desarrollo adecuado del potencial estadístico.

- Desarrollar los programas computacionales para el manejo de la información estructural de las proteínas, necesarios para el desarrollo del potencial estadístico.

- Probar el potencial desarrollado en 76 grupos de confórmeros de distintas proteínas.

VI. Desarrollo experimental

1. Material y métodos

1.1. Archivos PDB

Son un tipo de archivos en los cuales se almacena la información estructural de las proteínas en el Protein Data Bank, normalmente estos datos provienen de estudios de difracción de rayos X o de resonancia magnética nuclear. Este tipo de archivos contiene las posiciones en el espacio de la mayoría de los átomos que constituyen una proteína, también contiene información acerca de la estructura primaria de la proteína, el método usado en la determinación de su estructura, el organismo del cual proviene, entre otras [20] (Figura 8).

1.2. Lenguaje Python

Todos los programas empleados en el desarrollo del potencial estadístico fueron programados en lenguaje Python. Este es un lenguaje de programación interpretado, orientado a objetos y dinámico. Se considera a Python un lenguaje multiparadigma ya que en este lenguaje se pueden mezclar la programación imperativa, la programación funcional y la programación orientada a objetos. Python tiene una sintaxis muy limpia que favorece un código legible, además está disponible en multitud de plataformas (UNIX, Solaris, Linux, DOS, Windows, OS/2, Mac OS, etc.). Las características de Python permiten una programación modular, en la que cada módulo es un programa independiente que realiza una tarea específica [36].

```
HEADER    HYDROLASE(ZYMOGEN)                      16-JAN-87   2CGA
TITLE     BOVINE CHYMOTRYPSINOGEN A. X-RAY CRYSTAL STRUCTURE ANALYSIS
COMPND   2 MOLECULE: CHYMOTRYPSINOGEN A;
SOURCE   2 ORGANISM_SCIENTIFIC: BOS TAURUS;
AUTHOR    D.WANG,W.BODE,R.HUBER
>>>>>>
>>>>>>
JRNL        AUTH    D.WANG,W.BODE,R.HUBER
JRNL        TITL    BOVINE CHYMOTRYPSINOGEN A X-RAY CRYSTAL STRUCTURE
>>>>>>
REMARK   1
REMARK   1 REFERENCE 1
>>>>>>
SEQRES   1 A  245  CYS GLY VAL PRO ALA ILE GLN PRO VAL LEU SER GLY LEU
SEQRES   2 A  245  SER ARG ILE VAL ASN GLY GLU GLU ALA VAL PRO GLY SER
SEQRES   3 A  245  TRP PRO TRP GLN VAL SER LEU GLN ASP LYS THR GLY PHE
>>>>>>
>>>>>>
SEQRES  17 B  245  LEU VAL GLY ILE VAL SER TRP GLY SER SER THR CYS SER
SEQRES  18 B  245  THR SER THR PRO GLY VAL TYR ALA ARG VAL THR ALA LEU
SEQRES  19 B  245  VAL ASN TRP VAL GLN GLN THR LEU ALA ALA ASN
CRYST1   59.300   77.100  110.100  90.00  90.00  90.00 P 21 21 21    8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.016863  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012970  0.000000        0.00000
SCALE3      0.000000  0.000000  0.009083        0.00000
MTRIX1   1  0.987700  0.155000  0.017700        6.21700    1
MTRIX2   1  0.022800 -0.031400 -0.999200      115.61600    1
MTRIX3   1 -0.154300  0.987400 -0.034600       -3.74800    1
ATOM      1  N   CYS A   1     -10.656  55.938  41.808  1.00 11.66           N
ATOM      2  CA  CYS A   1     -10.044  57.246  41.343  1.00 11.66           C
ATOM      3  C   CYS A   1     -10.076  58.323  42.431  1.00 11.66           C
ATOM      4  O   CYS A   1     -10.772  58.097  43.448  1.00 11.66           O
ATOM      5  CB  CYS A   1     -10.807  57.718  40.066  1.00 11.66           C
>>>>>>
>>>>>>
>>>>>>
ATOM    744  N   ASN A 100     -13.152  77.724  22.378  1.00  8.65           N
ATOM    745  CA  ASN A 100     -14.213  76.940  23.011  1.00  8.65           C
ATOM    746  C   ASN A 100     -14.134  75.441  22.693  1.00  8.65           C
ATOM    747  O   ASN A 100     -13.706  75.062  21.563  1.00  8.65           O
>>>>>>
>>>>>>
ATOM   1461  N   VAL A 200      -9.212  70.793  39.923  1.00  9.30           N
ATOM   1462  CA  VAL A 200      -9.875  69.689  40.639  1.00  9.30           C
ATOM   1463  C   VAL A 200     -10.634  70.148  41.868  1.00  9.30           C
ATOM   1464  O   VAL A 200     -10.151  70.985  42.657  1.00  9.30           O
>>>>>>
HETATM 3601  O   HOH A 601     -20.008  66.224  26.138  1.00 26.69           O
HETATM 3602  O   HOH A 602     -21.333  66.182  28.756  1.00 18.10           O
HETATM 3603  O   HOH A 603     -18.000  68.022  22.774  1.00 34.03           O
MASTER      448    1    0   11   21    0    0    9 3927    2   20   38
END
```

Figura 8. Versión resumida de un archivo pdb representativo (2CGA). Los símbolos >>>>>> indican ausencia de varios renglones semejantes [20]

1.3. AMBER

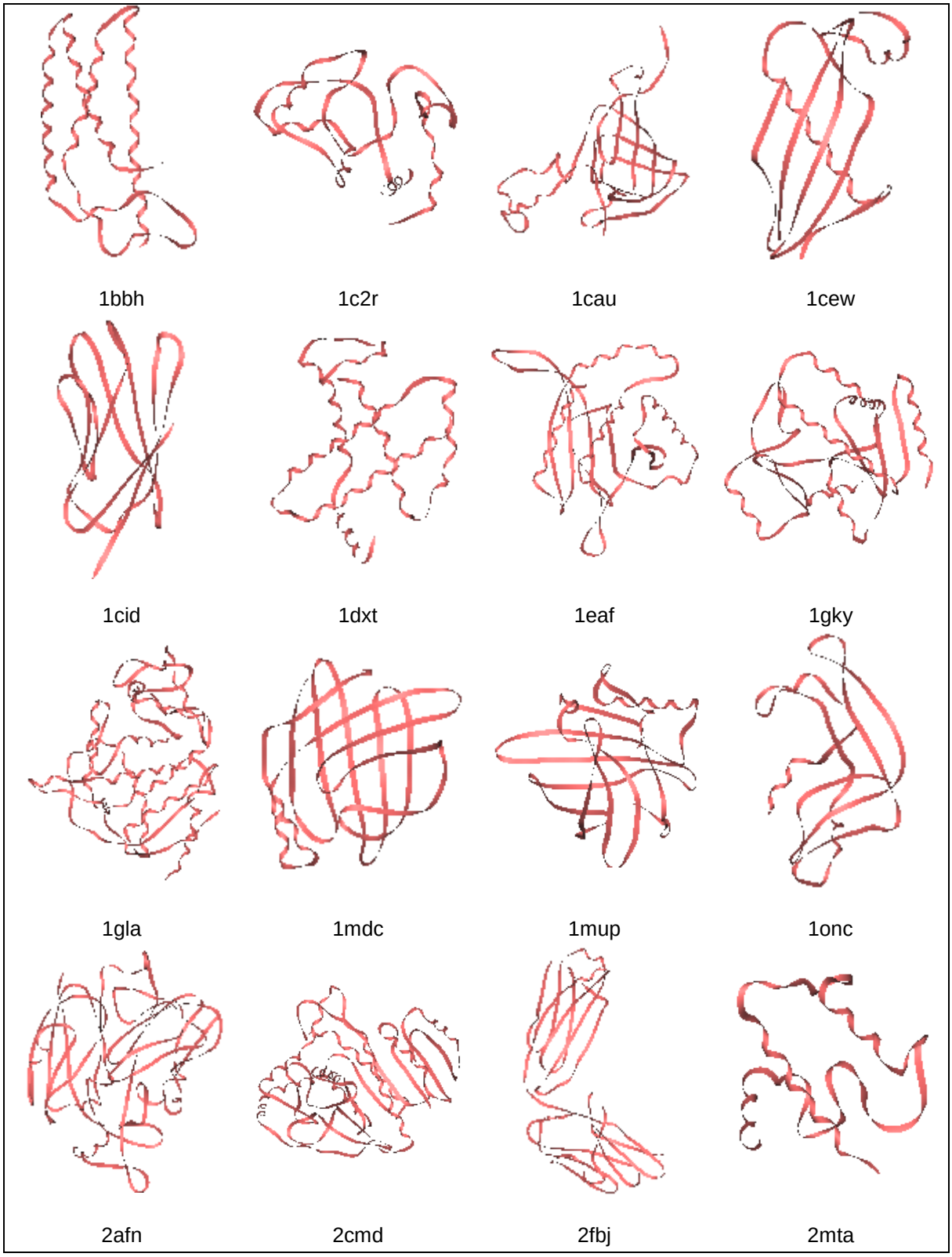
AMBER (Assisted Model Building with Energy Refinement) es el nombre colectivo de un conjunto de programas que permiten realizar y analizar simulaciones de dinámica molecular, particularmente para las proteínas, ácidos nucleicos y carbohidratos. Dentro de sus diversas funciones existe una llamada “protonate”, que se encarga de añadir hidrógenos a todos los átomos pesados en un archivo PDB[37]. Esta función de la versión 9 de AMBER se utilizó en la adición de hidrógenos a los archivos PDB empleados para la construcción del potencial.

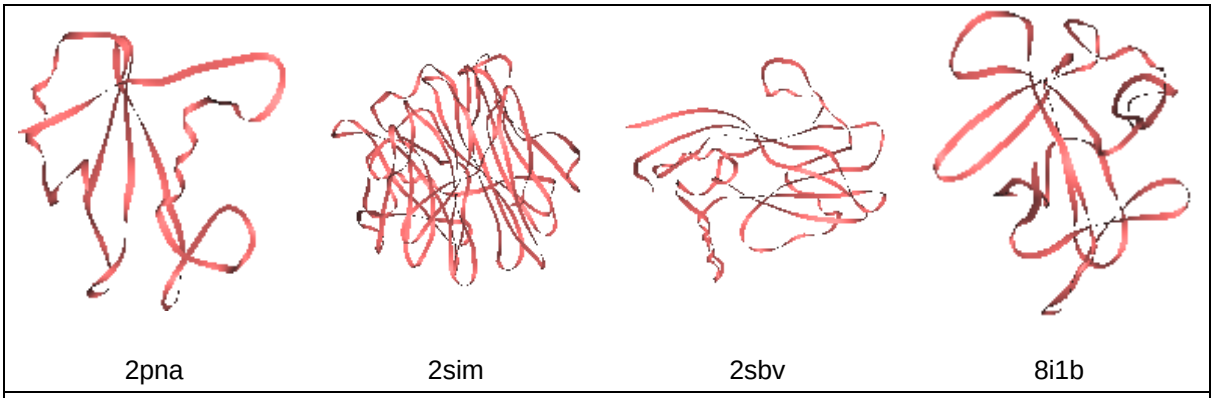
1.4. El conjunto de confórmeros

El desarrollo y evaluación de nuevas funciones de energía es crítico para el correcto modelado de las propiedades de macromoléculas biológicas Es por ello que las pruebas de discriminación de confórmeros, también llamados señuelos, se ha convertido en un enfoque ampliamente usado para probar y validar funciones de energía [38-39]. Los grupos de señuelos usados para probar el potencial desarrollado en este trabajo fueron los 20 generados por John y Sali [40] y los 56 generados por I-TASSER [41].

1.4.1. Señuelos de John y Sali (MOULDER)

Estos conjuntos de señuelos surgen de la selección de 20 secuencias de proteínas de entre 51 y 568 residuos, y con baja relación entre sí, para las cuales se construyeron 300 modelos por homología, considerando únicamente las posiciones de los átomos pesados y usando como plantilla para su creación la estructura más cercana a la proteína en cuestión. Estos modelos poseen al menos 5 residuos alineados de manera distinta.


1bbh
1c2r
1cau
1cew
1cid
1dxt
1eaf
1gky
1gla
1mdc
1mup
1onc
2afn
2cmd
2fbj
2mta

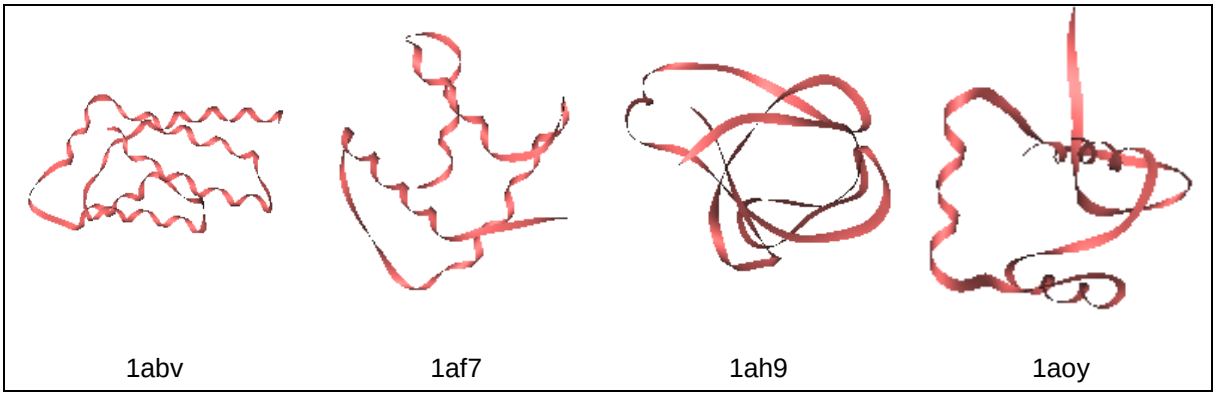


2pna 2sim 2sbv 8i1b

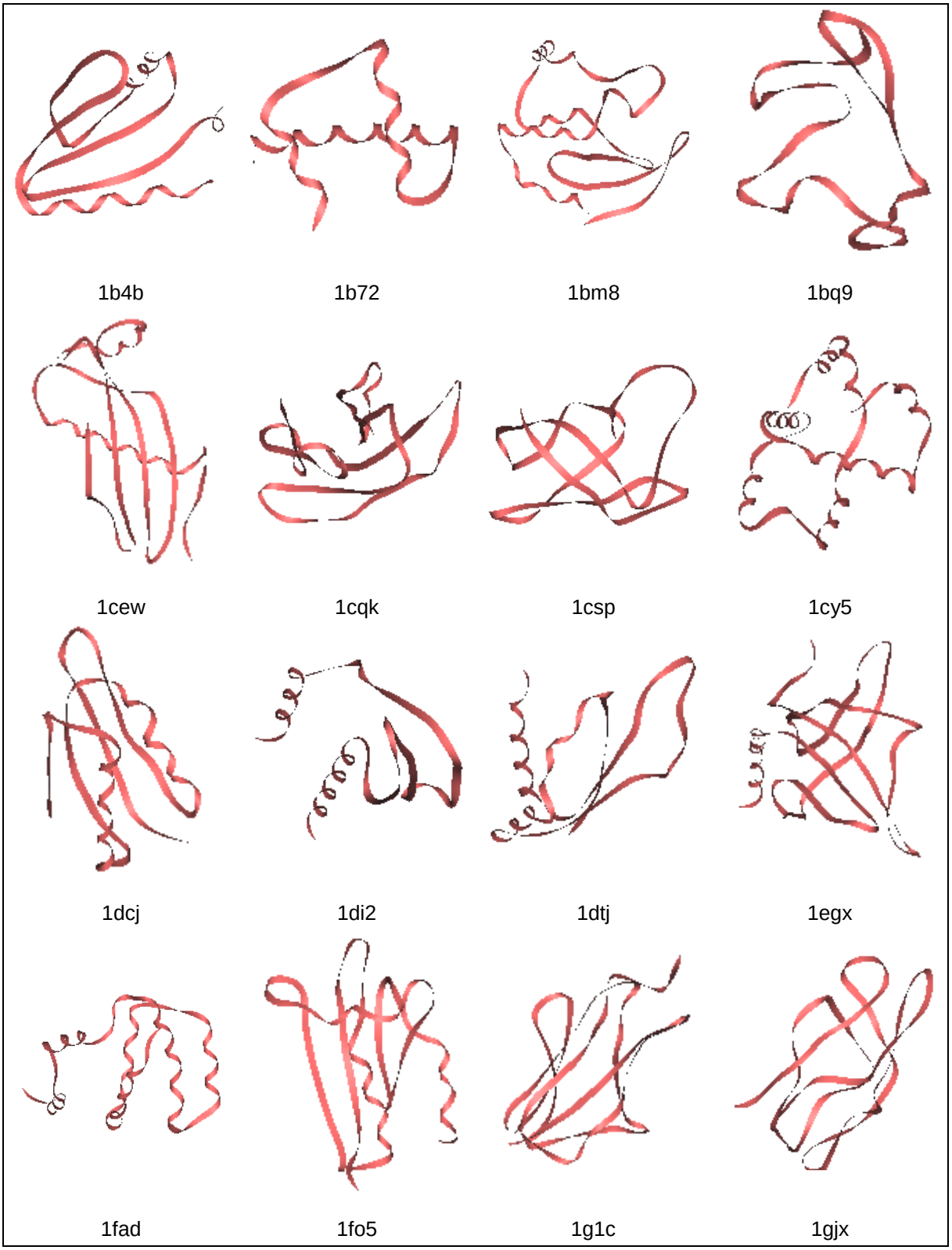
Tabla 2. Proteínas usadas como señuelos por John y Sali. Representación de su estructura en forma de listones, obtenidas mediante el software BetaMol [42]

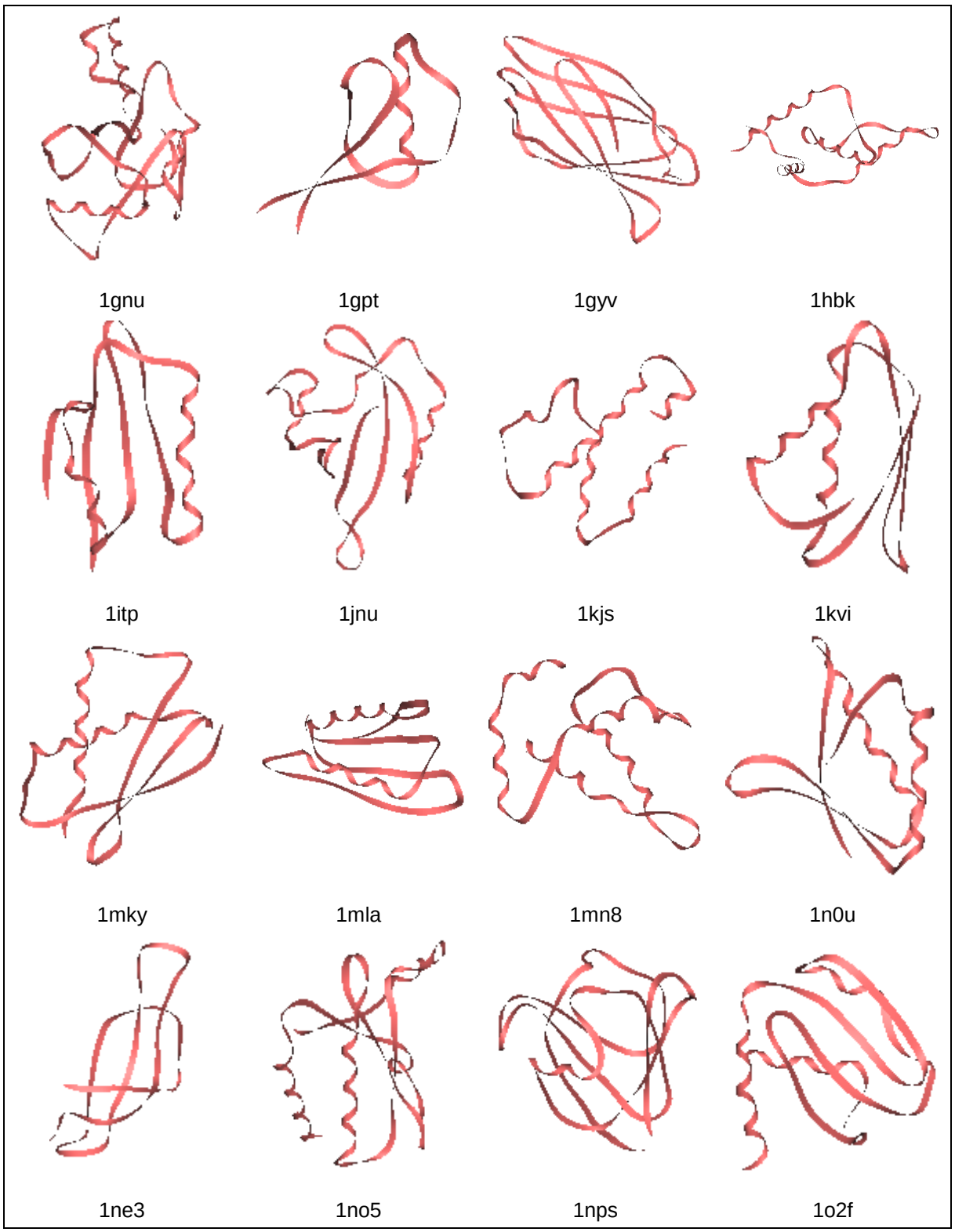
1.4.2. Señuelos de I-TASSER

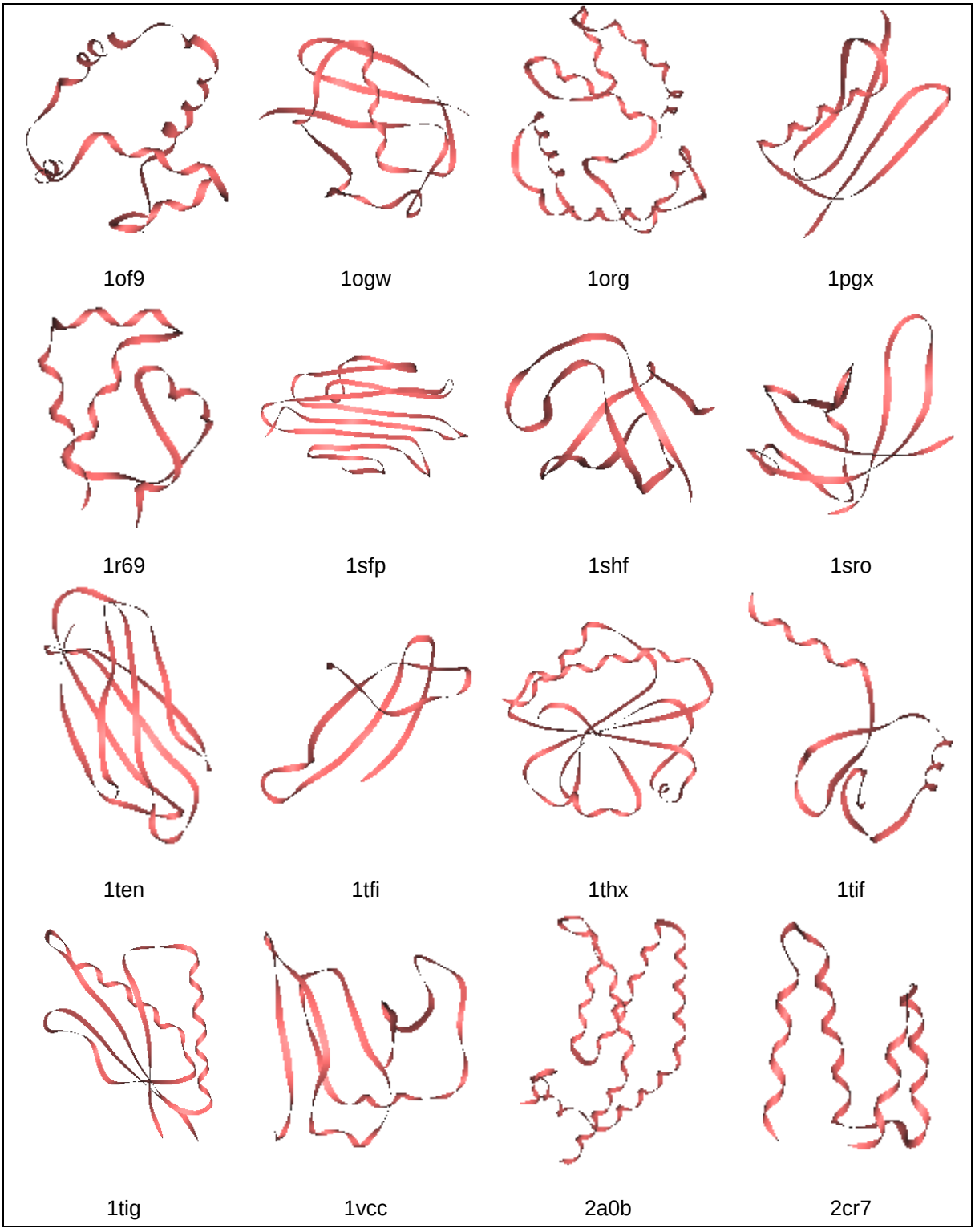
Estos conjuntos de señuelos surgen en base a 56 subunidades de proteínas no homólogas, para las cuales se diseñó la estructura del esqueleto por modelado *ab initio* y se generaron entre 12500 y 13200 modelos mediante simulaciones de mecánica molecular a baja temperatura. De entre todas las estructuras se seleccionaron por agrupamiento iterativo entre 300 y 500 confórmeros.



1abv 1af7 1ah9 1aoy


1b4b
1b72
1bm8
1bq9
1cew
1cqk
1csp
1cy5
1dcj
1di2
1dtj
1egx
1fad
1fo5
1g1c
1gjx


1gnu
1gpt
1gyv
1hbk
1itp
1jnu
1kjs
1kvi
1mky
1mla
1mn8
1n0u
1ne3
1no5
1nps
1o2f


1of9
1ogw
1org
1pgx
1r69
1sfp
1shf
1sro
1ten
1tfi
1thx
1tif
1tig
1vcc
2a0b
2cr7

2f3n 2pcy 2reb 256b

Tabla 3. Estructura de las proteínas usadas como señuelos por I-TASSER. Representación de su estructura en forma de listones, obtenidas mediante el software BetaMol [42]

2. Metodología

Se seleccionó una muestra de estructuras de proteínas del Protein Data Bank de acuerdo a las siguientes características:

- *Resueltas a través de difracción de rayos X.*
- *Resolución de la muestra de 2 Á o menos.*
- *Factor de refinamiento menor a 3.0*
- *Identidad de secuencia máxima del 30%.*
- *Sin residuos modificados*

Dadas estas especificaciones, se descargó una muestra de 2635 estructuras, las cuales posteriormente se sometieron a la función “protonate” de AMBER con la finalidad de añadir los átomos de hidrógeno a la estructura. Después de esto, se eliminaron 557 estructuras que no se procesaron correctamente, debido principalmente a la ausencia de las posiciones de varios residuos en el archivo PDB.

Para cada una de las 2078 estructuras completas se realizó el conteo de los puentes de hidrógeno formados por el esqueleto de las proteínas con base a dos parámetros: el ángulo donador-hidrógeno-aceptor (D-H-A) y la distancia hidrógeno-aceptor (H-A).

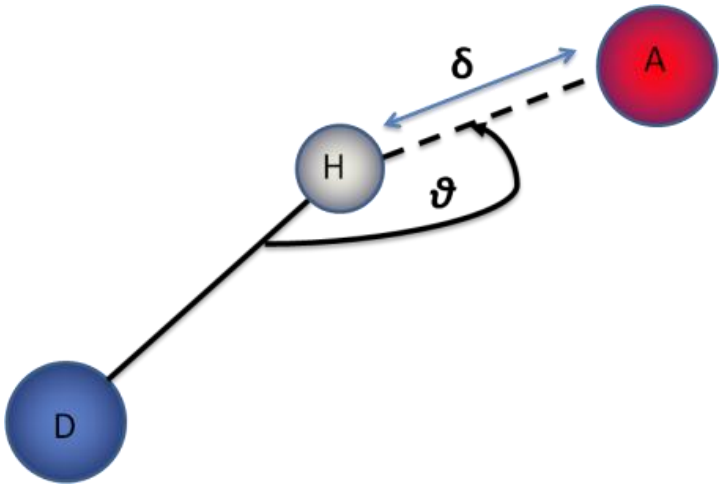


Figura 9. Representación de la parametrización elegida para caracterizar el puente de hidrógeno. D: donador (Nitrógeno), H: hidrógeno, A: Aceptor (Oxígeno)

La distancia H-A (δ) se consideró entre 1.6 y 2.5 Angstroms, divididos para su conteo de frecuencias en 9 intervalos iguales. El ángulo D-H-A (θ) se consideró para dos intervalos distintos, generando así dos potenciales, en el primer caso de 120 a 180 grados, cuyo potencial será identificado como PH1, y en el segundo de 90 a 180 grados, identificado como PH2; dichos intervalos fueron divididos para su conteo de frecuencias en 60 y 90 intervalos iguales, respectivamente.

La asignación de una energía para cada intervalo se realizó de una manera análoga a la distribución de Boltzmann, siguiendo el principio de Miyazawa y Jernigan, de la siguiente manera:

Distancia:

$$E_\delta = -ln\left(\frac{P_\delta}{P_\delta^r}\right)$$

Donde E_δ es la energía asociada a un intervalo de distancia, P_δ y P_δ^r son las probabilidades de encontrar un puente de hidrógeno en dicho intervalo, la primera de acuerdo a la muestra y la segunda considerando una distribución aleatoria.

Ángulo:

$$E_\theta = -ln\left(\frac{P_\theta}{P_\theta^r}\right)$$

Donde E_θ es la energía asociada a un intervalo de ángulo, P_θ P_δ^r son las probabilidades de encontrar un puente de hidrógeno en dicho intervalo, la primera de acuerdo a la muestra y la segunda considerando una distribución aleatoria.

De esta manera, el potencial definido para la geometría de un puente de hidrógeno formado por el esqueleto de una proteína (E_{PH}) queda definido como:

$$E_{PH} = E_{\delta} + E_{\theta}$$

O bien:

$$E_{PH} = -ln\left(\frac{P_{\delta}P_{\theta}}{P_{\delta}^{r}P_{\theta}^{r}}\right)$$

Finalmente, se añadieron los hidrógenos a las estructuras de los conjuntos de señuelos de MOULDER, y tanto en ellas como en los señuelos de I-TASSER, se probó el potencial generado asociando un valor de energía a cada confórmero dentro de un grupo de señuelos y determinando la posición correspondiente a la estructura nativa al ordenar los confórmeros en orden ascendiente de energías, tomando como una predicción acertada cuando la estructura nativa correspondía a la menor energía asociada, o posición número uno.

Para evaluar la calidad de la predicción se empleó un estadístico conocido como Z Score. Dicho estadístico representa el número de desviaciones estándar que un valor particular se aleja de la media de una muestra, en este caso en particular se define como:

$$Z\ Score = \frac{E_n - E_m}{s}$$

Donde E_m es el promedio de las energías asignadas a los señuelos de un grupo, s es su desviación estándar y E_n es la energía asignada a la estructura nativa.

Posteriormente se realizó una prueba t de Student para muestras relacionadas con la finalidad de comparar las medias de los Z Score obtenidos por los dos potenciales.

Por último, se compararon los resultados obtenidos por ambos potenciales con resultados obtenidos por otros potenciales probados en los mismos grupos de señuelos, tales como DOPE[43], DFIRE[44], RWplus [41], DBNI[45], OPUS-PSP[46], Multi_well[47] y DOKB[48], cuyas bases se describen brevemente a continuación:

- DOPE (Discrete Optimized Protein Energy) es un potencial distancia-dependiente derivado de la unión de las densidades de probabilidad de las coordenadas cartesianas de los átomos de una proteína y de las distancias entre ellos.

- DFIRE (Distance-scaled, Finite Ideal-gas Reference State) es un potencial de pares de contacto de todos los átomos de una proteína, que usa como estado de referencia una distribución uniforme de puntos en una esfera finita, semejante a la distribución de un gas ideal en un espacio finito.

- RWplus (Random walk) es un potencial de pares de contacto, dependiente de la orientación de las cadenas laterales de los residuos que constituyen una proteína, que usa como estado de referencia una cadena ideal generada por un movimiento aleatorio.

- DBNI (Delaunay-Based Nonlocal Interactions) es un potencial de pares de contacto basado en las interacciones entre 167 tipos de átomos, separados entre sí por más de 5 aminoácidos, y determinadas en base a la triangulación de Delauney.

- OPUS-PSP es un potencial de pares de contacto, dependiente de la orientación de los aminoácidos en contacto, representados como bloques.

- Multi_well es un potencial de pares de contacto entre los átomos presentes en la estructura secundaría de una proteína que pretende identificar la topología nativa de dicha estructura entre todas las posibilidades existentes.

- DOKB (Distance and Orientation dependent energy function of amino acid Key Blocks) es un potencial de pares de contacto, dependiente de la distancia y orientación de los aminoácidos en contacto, representados como bloques.

VII. Resultados

1. Conteo de puentes de hidrógeno

Bajo los parámetros establecidos para el desarrollo del potencial PH1, fueron contados 484,380 puentes de hidrógeno, en la distribución de dichos puentes se puede observar una frecuencia máxima para un ángulo N-H-O alrededor de los 159° (Figura 10), y una distancia H-O alrededor de los 2.0 Å (Figura 11).

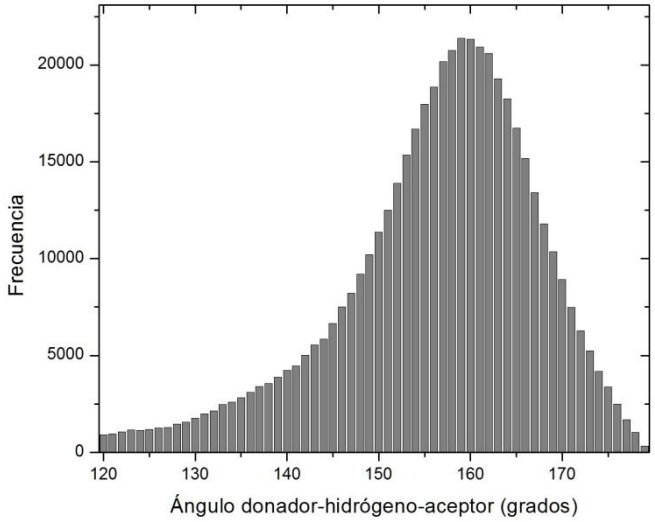


Figura 10. Distribución de frecuencias del ángulo D-H-A en los puentes de hidrógeno formados por el esqueleto de proteínas para PH1.

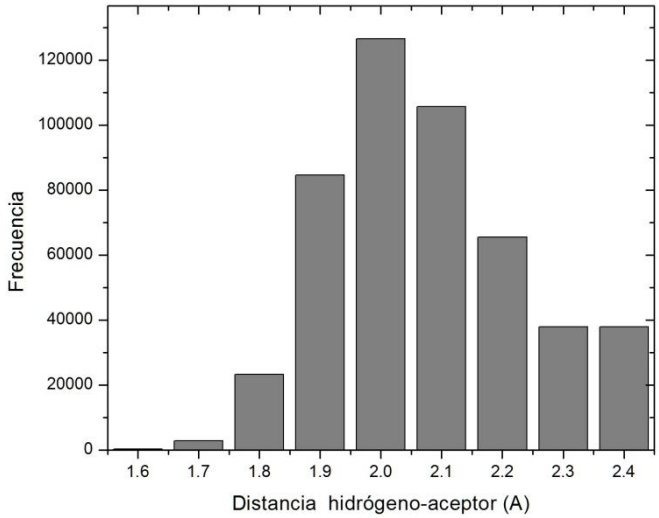


Figura 11. Distribución de frecuencias de la distancia H-A en los puentes de hidrógeno formados por el esqueleto de proteínas para PH1.

Por otro lado, bajo los parámetros establecidos en la generación del potencial PH2, se obtuvo un conteo total de 512,265 puentes de hidrógeno, siendo evidente que al extender el ángulo N-H-O hasta los 90°, surge la presencia de un máximo local alrededor de los 98° (Figura 12) y que esos puentes de hidrógeno corresponden principalmente a las distancias mayores en el intervalo estudiado, generando un máximo local cerca de los 2.4 Å (Figura 13).

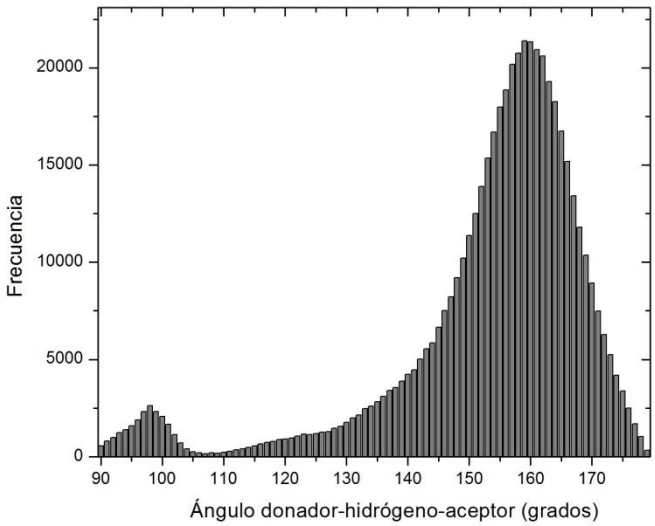


Figura 12. Distribución de frecuencias del ángulo D-H-A en los puentes de hidrógeno formados por el esqueleto de proteínas para PH2.

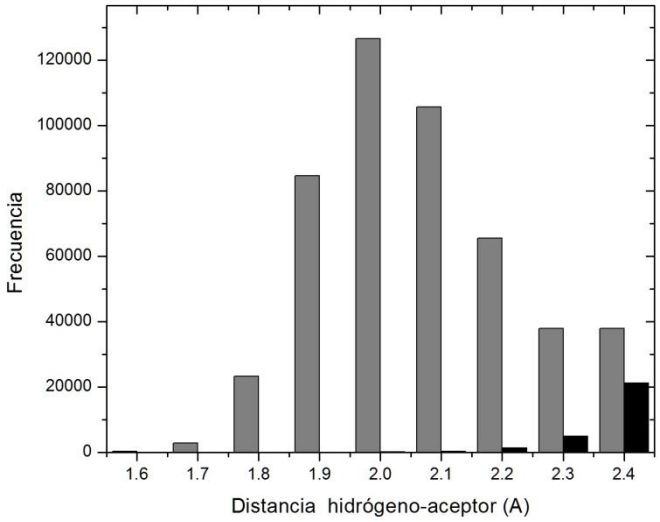


Figura 13.Distribución de frecuencias de la distancia H-A en los puentes de hidrógeno formados por el esqueleto de proteínas. Gris: PH1. Negro: PH2.

2. Construcción de los potenciales PH1 y PH2

La energía asociada a cada intervalo de ángulos y distancias se presentan respectivamente en las figuras 14 y 15 para el potencial PH1 y en las figuras 16 y 17 para el potencial PH2, observándose, por la naturaleza del potencial, mínimos para los ángulos y distancias con frecuencias máximas en el conteo de los puentes de hidrógeno.

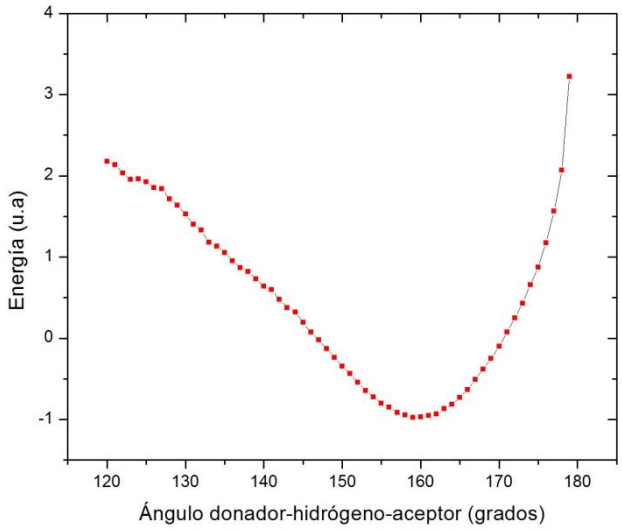


Figura 14.Componente angular de energía para puentes de hidrógeno formados por el esqueleto de proteínas para PH1.

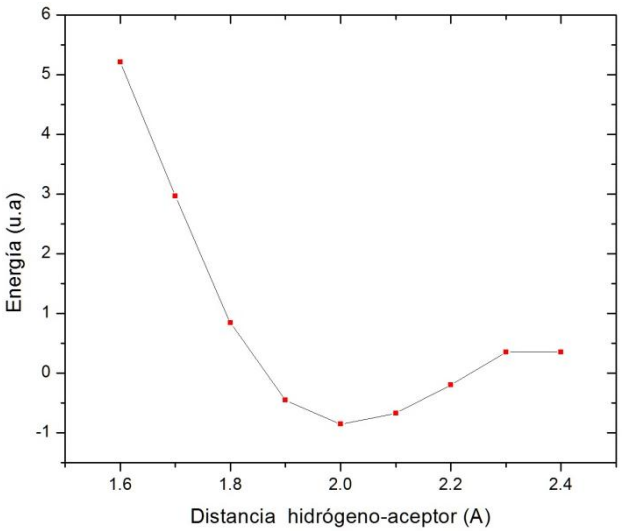


Figura 15. Componente longitudinal de energía para puentes de hidrógeno formados por el esqueleto de proteínas para PH1.

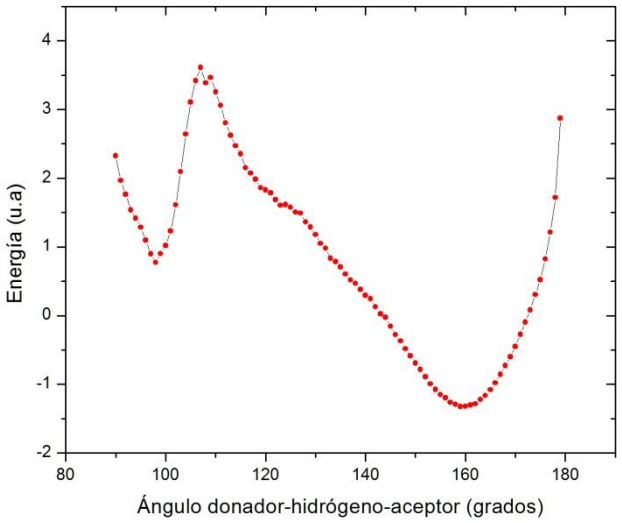


Figura 16. Componente angular de energía para puentes de hidrógeno formados por el esqueleto de proteínas para PH2.

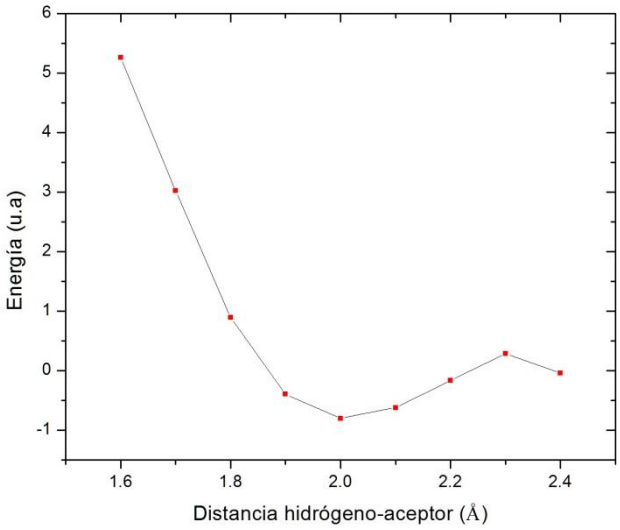


Figura 17. Componente longitudinal de energía para puentes de hidrógeno formados por el esqueleto de proteínas para PH2.

3. Prueba del potencial en los señuelos de MOULDER e I-TASSER

En seguida se presentan los resultados obtenidos al probar el potencial generado en los grupos de señuelos de MOULDER (Tabla 4) e I-TASSER (Tabla 5).

	PH1		PH2	
Proteína	**Posición**	**Z Score**	**Posición**	**Z Score**
1bbh	1	-2.74907525	1	-2.5530265
1c2r	1	-3.61485891	1	-3.82313237
1cau	41	-1.18749034	9	-1.83485037
1cew	123	-0.34506331	32	-1.23962819
1cid	195	0.51626809	96	-0.49428928
1dxt	1	-2.39552338	1	-2.20219094
1eaf	1	-3.73541196	1	-3.77880232
1gky	1	-7.37009102	1	-7.45661542
1lga	1	-3.1614881	1	-2.91995013
1mdc	1	-2.97470612	1	-2.74116719
1mup	71	-0.76670915	38	-1.20736458
1onc	2	-2.66940416	1	-2.75202155
2afn	1	-4.26399888	1	-4.51964841
2cmd	1	-4.94161792	1	-4.54217852
2fbj	1	-4.57343534	1	-4.40707386
2mta	1	-2.61413732	1	-2.83861731
2pna	298	2.16910781	290	1.73669518
2sim	1	-6.96207069	1	-5.85884514
4sbv	265	1.18755385	55	-0.97325987
8i1b	8	-1.78674661	1	-2.24662009
Z Score promedio		**-2.61194493**		**-2.83262934**
Predicciones acertadas		**12/20**		**14/20**

Tabla 4. Evaluación de los grupos de señuelos de MOULDER

	PH1		PH2	
Proteína	**Posición**	**Z Score**	**Posición**	**Z Score**
1abv	344	0.30523797	249	-0.08762278
1af7	1	-3.60315801	1	-3.63936077
1ah9	384	0.67537897	191	-0.44591902
1aoy	2	-2.79544947	2	-3.11725445
1b4b	1	-3.22468729	1	-4.0223333
1b72	1	-4.90663021	1	-5.52232743
1bm8	1	-5.83275499	1	-7.89892943
1bq9	1	-5.36972688	1	-6.8168403
1cew	25	-1.53791769	1	-2.64521962
1cqk	1	-3.43526075	1	-4.43503288
1csp	1	-3.96759139	1	-4.58750175
1cy5	1	-4.23195498	1	-4.34117909
1dcj	294	0.08500053	150	-0.55800329
1di2	1	-4.74652154	1	-4.86619158
1dtj	1	-3.06102249	1	-3.44334233
1egx	354	2.80968574	353	2.43445335
1fad	139	-0.59340136	86	-0.91422469
1fo5	1	-3.29440444	1	-3.42036595
1g1c	1	-4.39011172	1	-5.39355264
1gjx	527	5.74307941	527	5.37210171
1gnu	1	-6.49389257	1	-7.40885479
1gpt	272	0.21813846	171	-0.35188383
1gyv	1	-7.02674855	1	-7.78514056
1hbk	1	-3.65937255	1	-3.73246587
1itp	125	-0.70037111	27	-1.58978507
1jnu	1	-3.19382682	1	-3.86697451
1kjs	134	-0.76079596	210	-0.23933625
1kvi	534	2.04442027	528	1.69023347
1mky	1	-5.48249767	1	-6.16116702
1mla	1	-6.4952989	1	-7.35763892
1mn8	1	-7.08903206	1	-8.02973286
1n0u	1	-3.84295038	1	-4.34009467
1ne3	568	5.48915571	568	4.93054728
1no5	1	-4.44133847	1	-5.05493457
1nps	1	-5.22434153	1	-6.5053715
1o2f	1	-2.86953758	1	-3.50188829
1of9	509	2.68273649	509	3.12391875
1ogw	1	-6.80320238	1	-7.78459198
1org	1	-5.51205321	1	-5.71311724
1pgx	236	-5.77511146	432	-6.84394427
1r69	12	-1.77226912	9	-1.86276312
1sfp	1	-5.98101545	1	-6.85099922
1shf	1	-5.03743164	1	-6.28009683

1sro	33	-1.49143752	18	-1.75272155
1ten	1	-5.83654064	1	-6.49394049
1tfi	6	-2.24740849	2	-3.04691272
1thx	1	-6.32992299	1	-6.79020965
1tif	1	-6.74094449	1	-7.96304933
1tig	1	-5.4232322	1	-5.98595996
1vcc	1	-5.18961014	1	-6.50632435
256b	1	-5.62513586	1	-5.63334765
2a0b	1	-5.65071192	1	-5.86198241
2cr7	16	-2.04752223	16	-2.02790022
2f3n	1	-4.16125455	1	-4.57773549
2pcy	1	-3.89055997	1	-5.08778266
2reb	1	-3.93593287	1	-4.33719476
Z Score promedio		**-3.24117188**		**-3.8511649**
Predicciones acertadas		**38/56**		**39/56**
Tabla 5. Evaluación de los grupos de señuelos de I-TASSER				

La Tabla 6 muestra el número de predicciones acertadas por distintos potenciales en la evaluación de los señuelos de MOULDER e I-TASSER. Debido a que la prueba t de Student para muestras relacionadas demostró que existe una diferencia significativa entre los potenciales PH1 y PH2, solo se incluyó el potencial PH2 en esta comparación ya que este presentó un mayor Z score promedio, así como un mayor número de predicciones acertadas.

	Señuelos	
Potencial	**MOULDER**	**I-TASSER**
DOPE	19 (-3.09)	30 (-2.18)
DFIRE	19 (-2.79)	47 (-3.58)
RWplus	19 (-3.04)	56 (-5.38)
DBNI	19 (-3.99)	42 (-3.63)
OPUS-PSP	19 (n.d.)	45 (n.d)
Multi_well	19 (n.d.)	16 (n.d)
DOKB	19 (n.d.)	53 (n.d)
PH2	14 (-2.83)	39 (-3.85)

Tabla 6. Desempeño de distintos potenciales en los grupos de señuelos MOULDER e I-TASSER. Entre paréntesis se muestra el Z Score promedio obtenido por cada potencial.

VIII. Discusión de resultados

El desarrollo de funciones energéticas y campos de fuerza para estudiar el comportamiento de sistemas moleculares es uno de los objetivos principales en el área de fisicoquímica. La predicción de la estructura nativa de proteínas en base a su secuencia de aminoácidos, la simulación del proceso de plegado y el cálculo de su estabilidad, se encuentran entre las metas más ambiciosas de la investigación contemporánea en la teoría biomolecular, para así entender diversos aspectos de su función y poder emplear dicho conocimiento en el diseño de nuevos fármacos [1-2, 49].

Una de las tareas más retadoras en la predicción de la estructura terciaria de proteínas es la de distinguir la conformación nativa de una proteína de entre un grupo de señuelos con conformación similar, encargados de dicha tarea se encuentran los campos de fuerza y los potenciales basados en el conocimiento, los cuales han tenido diversas aplicaciones en el diseño de proteínas y el acoplamiento molecular de ellas [48].

En el estudio aquí presentado se generaron dos potenciales estadísticos basados en la geometría del puente de hidrógeno en proteínas y se evaluó su desempeño frente a 76 conjuntos de señuelos.

En la generación de los potenciales PH1 y PH2 se encontró que la mayor cantidad de puentes de hidrógeno presentes en la cadena principal de proteínas se encuentra alrededor de 2.0 Å para la distancia H-O y 159° para el ángulo N-H-O, valores aproximados a los reportados por Baker y Hubbard [24], los cuales son 2.05 Å y 155° respectivamente, la ligera diferencia existente puede deberse principalmente al tamaño de la muestra empleada, ya que en el texto citado se estudiaron sólo 15 proteínas diferentes mientras que en el presente

trabajo se analizaron 2078, esta puede ser también la razón de que en dicho estudio no se encontrara el máximo local de frecuencias alrededor de los 98° para el ángulo N-H-O hallado en la construcción del potencial PH2. Debido a la manera en que los potenciales PH1 y PH2 se construyeron, el comportamiento descrito para las frecuencias angulares y longitudinales se puede observar de manera inversa en los componentes energéticos correspondientes, es decir, con la presencia de mínimos de energía para los intervalos de frecuencia máxima.

De acuerdo con los resultados obtenidos por la prueba t de Student para muestras relacionadas se puede establecer, con un 95% de confianza, que el potencial PH2 posee un mayor poder predictivo en relación al potencial PH1, con un valor aproximado de 70%, esto sugiere, con relación a la construcción de dichos potenciales, que el conteo de puentes de hidrógeno formados por el esqueleto de proteínas, en el intervalo [90° - 120°], es un factor importante para la discriminación de la estructura nativa de una proteína en un conjunto de señuelos.

Para evaluar el desempeño del potencial PH2 al ser probado en los señuelos de MOULDER e I-TASSER, se comparó con otros potenciales cuya evaluación en los mismos grupos de señuelos había sido previamente reportada [41,43-48]. Comparando el número de predicciones correctas y el Z Score promedio, se puede observar que, para los señuelos de MOULDER, PH2 obtiene el peor resultado, mientras que para los señuelos de I-TASSER se supera en el número de predicciones correctas únicamente a los potenciales Multi_well y DOPE pero se obtiene el segundo mejor valor en cuanto al promedio de los Z Score, solo por debajo de RWplus. Este hallazgo es congruente, considerando el hecho de que los potenciales usados para la comparación fueron construidos en base a interacciones entre todos los tipos de átomos presentes en una proteína, los cuales pueden llegar a ser hasta 167 [45], por lo que el

nivel de descripción que proporcionan debiera ser mucho mayor que el de un potencial construido en base a la interacción de sólo 3 tipos de átomo, que por la naturaleza de su construcción, solo podría discriminar entre confórmeros en base a la cantidad de estructura secundaria que presenten, pero no podría diferenciar cambios conformacionales en secciones sin estructura secundaria de una proteína ni modificaciones en las cadenas laterales de los aminoácidos. Tomando en cuenta estos aspectos, es de resaltar el hecho de que, a pesar de que PH2 no obtiene el mayor número de predicciones correctas, el Z Score promedio se encuentre solo por debajo del obtenido por RWplus, lo cual refleja que las estructuras predichas correctamente por PH2 se encuentran ampliamente diferenciadas del resto de los confórmeros.

IX. Conclusiones

En el presente trabajo se mostró que la capacidad de un potencial estadístico basado en dos descriptores, uno de distancia y otro de direccionalidad, para los puentes de hidrógeno formados por la cadena principal de proteínas con estructura terciaria conocida, se encuentra alrededor del 70% para predecir la estructura nativa de una proteína en un conjunto de señuelos. Encontrando que los puentes de hidrógeno con un ángulo entre 90° y 120° son un factor importante para la realización de dicha discriminación.

X. Propuestas

En el presente trabajo se desarrolló un potencial estadístico basados en distintos parámetros geométricos de los puentes de hidrógeno formados por la cadena principal de las proteínas, determinándose que dicho potencial puede predecir la estructura nativa de una proteína en un conjunto de señuelos en el 70% de los casos.

Lo expuesto anteriormente permite pensar que el estudio más detallado de los puentes de hidrógeno podría incrementar de manera considerable el poder predictivo del potencial. Para la realización de dicho estudio se propone la consideración de dos descriptores más de direccionalidad, los cuales involucran la medición de los ángulos axial y ecuatorial formados por un hidrógeno y el plano encontrado en el enlace peptídico.

Por otra parte, la limitación del potencial para distinguir los cambios conformacionales de residuos no participantes en la formación de estructura secundaria se podría superar incluyendo el estudio de otros tipos de átomos participantes en la formación de puentes de hidrógeno, lo cual tomaría en consideración la ocurrencia de estas interacciones entre la cadena principal y las cadenas laterales de los aminoácidos, así como entre cadenas laterales.

XI. Referencias

[1] Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol*. 2000; 18(3): 283-287

[2] Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001; 294(5540): 93-96

[3] EBI [sede Web]. Hinxton, Cambridgeshire, United Kingdom: UniProt; 2013 [acceso 30 de Agosto de 2013]. Current Release Statistics. Disponible en: http://www.ebi.ac.uk/uniprot/TrEMBLstats

[4] RCSB PDB [sede Web]. RCSB. 2013 [acceso 30 de Agosto de 2013]. Disponible en: http://www.rcsb.org/pdb/home/home.do

[5] Murzin AG. Progress in protein structure prediction. *Nat. Struct. Biol*. 2001; 8(2): 110-112

[6] Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature*. 1992; 358(6381): 86-89

[7] Vingron M, Waterman MS, Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol*. 1994; 235(1): 1-12

[8] Karplus K, Barett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 1998; 14(10): 846-856

[9] Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 1999; 287(4): 797-815

[10] David R, Korenberg MJ, Hunter IW. 3D-1D threading methods for protein fold recognition. *Pharmacogenomics.* 2000; 1(4): 445-455

[11] Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* 2001; 10(11): 2354-2362

[12] Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci.* U.S.A. 2004; 101(20): 7594-7599

[13] Anfinsen CB. Principles that Govern the Folding of Protein Chains. *Science.* 1973; 181(4096): 223-230

[14] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations*. J. Comput. Chem.* 1983; 4(2): 187-217

[15] Halaren TA. Potential energy functions. *Curr. Opin. Struct. Biol.* 1995; 5(2): 205-210

[16] Mackerell AD Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* 2004; 25(13): 1584-1604

[17] Gō N. Theoretical Studies of Protein Holding. *Annu. Rev. Biophys. Bioeng.* 1983; 12: 183-210

[18] Dehouck Y, Bilis D, Rooman M. A New Generation of Statistical Potentials for Proteins. *Biophys. J.* 2006; 90(11): 4010-4017

[19] Hubbard RE, Kamran M. Hydrogen Bonds in Proteins: Role and Strength. Encyclopedia of Life Sciences (ELS), John Wiley & Sons. 2010: 1-7

[20] Whitford D. Proteins: structure and function. Chichester, West Sussex, England. John Wiley & Sons Ltd. 2005

[21] Kessel A, Ben-Tal N. Introduction to proteins: structure, function and motion. Raton, Florida, United States of America. CRC Press (Taylor and Francis LLC). 2011

[22] Müller-Ester W (Versión española por Centelles JJ). Bioquímica: Fundamentos para medicina y ciencias de la vida. Barcelona, España. Reverté. 2008

[23] Karshikoff A. Non-covalent interactions in proteins. London, England. Imperial College Press. 2006

[24] Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. *Prog. Biophys. Mo.l Biol.* 1984; 44(2): 97-179

[25] Bartlett AI, Radford SE. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat. Struct. Mol. Biol.* 2009; 16(6): 582-588

[26] Brockwell DJ, Radford SE. Intermediates: ubiquitous species on folding energy landscapes? *Curr. Opin. Struct. Biol*. 2007; 17(1): 30-37

[27] Onuchic JN, Wolynes PG. Theory of protein folding. *Curr. Opin. Struct. Biol.* 2004; 14(1): 70-75

[28] Hartl FU, Hayer-Hartl M. Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* 2009; 16(6): 574-581

[29] Wang ZX. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* 1998; 11(8): 621-626

[30] Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Curr. Opin. Struct. Biol.* 2002; 12(2): 176-181

[31] Garduño-Juárez R, Morales LB. A Genetic Algorithm with Conformational Memories for Structure Prediction of Polypeptides. *J. Biomol. Struct. & Dyn.* 2003; 21(1): 41-63

[32] Miyazawa S, Jernigan RL.; Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules.* 1985; 18(3): 534–552

[33] Sippl MJ.; Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins, *J. Mol. Biol.* 1990; 213(4): 859-883

[34] Krishnamoorthy B, Tropsha A.; Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations, *Bioinformatics.* 2003; 19(12): 1540–1548

[35] Sánchez-González G, Kim JK, Kim DS, Garduño-Juárez R. A beta-complex statistical four body contact potential combined with a hydrogen bond statistical potential recognizes the correct native structure from protein decoy sets. *Proteins*. 2013; 81(8): 1420–1433

[36] Lee W, Kim HY. Genetic algorithm implementation in Python. Computer and Information Science. Fourth Annual ACIS International Conference. 2005: 8- 11

[37] Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. *J. Comput. Chem*. 2005; 26(16): 1668-1688

[38] Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins.* 2002; 48(2): 404-422

[39] Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 1997; 266(4): 831-846

[40] John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* 2003; 31(14): 3982-3992

[41] Zhang J, Zhang Y. A Distance-Dependent Atomic Potential Derived from Random-Walk Ideal Chain Reference State for Protein Fold Selection and Structure Prediction. *PLoS One.* 2010: 5(10): e15386

[42] Cho Y, Kim JK, Ryu J, Won CI, Kim CM, Kim D, Kim DS. BetaMol: a molecular modeling, analysis and visualization software based on the beta-complex and the quasi-triangulation, *J. Adv. Mech. Des. Syst. Manuf.* 2012: 6(3): 389-403

[43] Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 2006; 15(11): 2507–2524

[44] Zhou HY, Zhou YQ (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 2002; 11(11): 2714–2726

[45] Mirzaie M, Sadeghi M. Delaunay-based nonlocal interactions are sufficient and accurate in protein fold recognition. *Proteins.* 2013: doi: 10.1002/prot.24407 [Artículo aceptado]

[46] Ma J. Explicit orientation dependence in empirical potentials and its significance to side-chain modeling. *Acc. Chem. Res.* 2009; 42(8): 1087-96

[47] Sun W, He J. Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topologies. *Proteins.* 2009; 77(1): 159-173

[48] Chen L, He J. A distance and orientation dependent energy function of amino acid key blocks. *Biopolymers*. 2013: doi: 10.1002/bip.22440 [Artículo aceptado]

[49] Sippl MJ. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 1995; 5(2): 229-235