TESIS DEFENDIDA POR

Rosario Ivetth Corona de la Fuente

Y APROBADA POR EL SIGUIENTE COMITÉ

Dr. Carlos Alberto Brizuela Rodríguez

Director del Comité

Dr. José Alberto Fernández Zepeda

Miembro del Comité

Dr. Joaquín Álvarez Gallegos

Miembro del Comité

Dr. Alejandro Martínez Ruiz

Miembro del Comité

Dr. Hugo Homero Hidalgo Silva

Coordinador del programa de posgrado en Ciencias de la Computación

Dr. David Hilario Covarrubias Rosales

Director de Estudios de Posgrado

9 de diciembre de 2010

CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA

PROGRAMA DE POSGRADO EN CIENCIAS EN CIENCIAS DE LA COMPUTACIÓN

ANÁLISIS COMPARATIVO DE DOS HEURÍSTICAS PARA EL PROBLEMA DE EMPAQUETAMIENTO DE LA CADENA LATERAL EN PROTEÍNAS

TESIS

que para cubrir parcialmente los requisitos necesarios para obtener el grado de

MAESTRO EN CIENCIAS

Presenta:

ROSARIO IVETTH CORONA DE LA FUENTE

Ensenada, Baja California, México, diciembre de 2010

RESUMEN de la tesis de **ROSARIO IVETTH CORONA DE LA FUENTE**, presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en CIENCIAS DE LA COMPUTACIÓN. Ensenada, Baja California, diciembre de 2010.

ANÁLISIS COMPARATIVO DE DOS HEURÍSTICAS PARA EL PROBLEMA DE EMPAQUETAMIENTO DE LA CADENA LATERAL EN PROTEÍNAS

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodríguez

Director de Tesis

Actualmente, es más fácil determinar la secuencia de aminoácidos de una proteína que la estructura de la misma, por lo que se intenta utilizar herramientas computacionales que predigan la estructura a partir de la secuencia de aminoácidos. A esto se le conoce como el problema de predicción de estructuras de proteínas, el cual es uno de los problemas sin resolver más importantes en biología molecular y biocomputación.

El problema de predicción de estructuras se puede dividir en varios subproblemas. Uno de ellos es el problema de empaquetamiento de la cadena lateral en proteínas (PSCPP, por sus siglas en inglés). Este problema consiste en determinar la conformación de la proteína conociendo, además de la secuencia de aminoácidos, la conformación de la columna vertebral de la misma.

En este trabajo se presenta un análisis experimental comparativo entre los métodos SCWRL4 y OPUS-Rota, que a la fecha son los que presentan las soluciones con mejor exactitud. Estos métodos se componen de tres elementos principales: un algoritmo de optimización, una función de energía que se desea minimizar, y una biblioteca de rotámeros.

Para hacer la comparación se propone un conjunto de casos de prueba con diferentes criterios que aseguran la buena calidad de los modelos del conjunto y, poder así, realizar una comparación justa. Uno de estos criterios está relacionado con la clase de estructura a la que pertenece la proteína siguiendo la clasificación conocida como SCOP. Además se relaciona el grupo al que pertenece cada proteína según una clasificación de propiedades enzimáticas de las mismas. Esta clasificación se denomina EC (*Enzyme Commission*).

Los métodos mencionados se han probado con anterioridad en casos específicos, bajo diferentes condiciones para cada algoritmo, por lo que es importante una comparación que permita evaluar el desempeño de ambos métodos en igualdad de circunstancias.

Los resultados experimentales muestran que ambos métodos tienen un desempeño similar, tanto en calidad de soluciones como en tiempo de ejecución. En ambos casos, la calidad de la solución es poco sensible al tipo de estructura según la clasificación SCOP. Tampoco mostraron sensibilidad en función del grupo EC al que pertenecían las proteínas. Sin embargo, ambos métodos se mostraron sensibles al tipo de aminoácido

a predecir. En este sentido es más fácil predecir la estructura de aminoácidos como fenilalanina e isoleucina que aminoácidos tales como serina y ácido glutámico.

El método OPUS-Rota, el cual está basado en recocido simulado, es más flexible en el sentido de que resulta fácil incluir otras funciones de energía y bibliotecas de rotámeros. Por esto se proponen algunas ideas para el diseño de un método de predicción de estructura basado en este algoritmo de optimización.

Palabras Clave: Empaquetamiento de la cadena lateral, estructura de proteínas, biblioteca de rotámeros, plegamiento de proteínas, casos de prueba, SCOP, EC.

ABSTRACT of the thesis presented by **ROSARIO IVETTH CORONA DE LA FUENTE**, in partial fulfillment of the requirements of the MASTER OF SCIENCE degree in COMPUTER SCIENCE. Ensenada, Baja California, december 2010.

COMPARATIVE ANALYSIS OF TWO HEURISTICS FOR THE PROTEIN SIDE CHAIN PACKING PROBLEM

With the current technology it is easier to determine the amino acid sequence of a protein than its structure, thus the importance of developing computational methods and tools to predict the structure from the amino acid sequence. This prediction problem is known as the protein structure prediction and is one of the most important open problems in molecular and computational biology.

The problem of structure prediction can be divided into several subproblems. One of them is the protein side-chain packing problem (PSCPP), which consists in determining the protein structure knowing the amino acid sequence, and the conformation of the backbone.

We present a comparative experimental analysis of the methods SCWRL4 and OPUS-Rota, which are to date, the ones with more accurate solutions. These methods consist of three main components: an optimization algorithm, an energy function that we want to minimize, and a rotamer library.

To make the comparison, a set of test instances is proposed based on different criteria, to ensure good sampling of all known structures and to make a fair comparison. One of these criteria is the class of structure to which it belongs, according to the classification known as SCOP. Another criterion to classify the proteins is related to the group to which each protein belongs according to the enzymatic properties it has. This classification is reffered to as EC (Enzyme Commission).

The methods have been previously tested on specific instances, under different conditions for each algorithm, therefore a comparison is important to evaluate the performance of both methods under the same conditions.

The results show that both methods have similar performance, in the accuracy of its solutions, and in the computation time. In both cases the quality of the solution is not sensitive to neither the type of structure according to the SCOP nor to the EC classification. However, both methods were sensitive to the type of amino acid. In this sense it is easier to predict the structure of amino acids such as phenylalanine and isoleucine than amino acids such as serine or glutamic acid.

The method OPUS-Rota, which is based on simulated annealing, is more flexible, in the sense that it is easy to include in this algorithm other energy functions and rotamer libraries. Thus, we propose some ideas for designing a structure prediction method based on this optimization algorithm.

Keywords: Side-chain packing, protein structure, rotamer library, protein folding, test instances, SCOP, EC.

A mi familia

Agradecimientos

A mi esposo, Rodrigo, quien ha sido un pilar importante en mi vida.

A mis padres, Heriberto y Rosalinda, quienes me han dado su apoyo para lograr mis sueños, pero más importante, porque siempre me permitieron soñar.

A mis hermanos, Rosalinda, Jesús Heriberto y Pedro Enrique quienes me han hecho saber que cuento con ellos.

Al Dr. Carlos Alberto Brizuela Rodríguez, mi director de tesis, y a los miembros de mi comité de tesis, por su apoyo y confianza.

A Daniel Eduardo y César Alberto, por compartir la experiencia de la maestría conmigo, y por las amenas discusiones que se generaban durante las horas de estudio.

A mis demás compañeros de generación, a mis compañeros del cubo 103 (Fermín, Vanessa, Daniel Fajardo, Ismael, Héctor, Daniel Brubeck y José), a los integrantes del grupo de biocomputación (Israel, Mauricio y Sarita) y a compañeros de otras generaciones que convivieron conmigo durante la realización de mi tesis y crearon un ambiente ameno facilitando las labores diarias.

A los demás investigadores y personal del departamento de Ciencias de la Computación del CICESE que de alguna u otra forma ayudaron a que mi estancia en la maestría fuera lo más productiva posible.

A todas las personas con las que conviví en UNCC, quienes me recibieron con las puertas abiertas y me mostraron su apoyo, haciendome sentir en casa.

Al CONACYT, por su apoyo ecnómico que fue fundamental en la realización de mis estudios de maestría.

Contenido

Página

Contenido (continuación)

Lista de Figuras

Lista de Figuras (continuación)

Lista de Figuras (continuación)

Lista de Tablas

Capítulo I

INTRODUCCIÓN

I.1 Antecedentes y motivación

La biocomputación es un área interdisciplinaria que incorpora conocimientos de biología, química, ciencias de la computación, estadística, termodinámica, entre otras, para resolver problemas de biología molecular. La biocomputación tiene varios objetos de estudio entre los que se encuentran el ADN (ácido desoxirribonucleico), el ARN (ácido ribonucleico) y las proteínas.

Se sabe que el ADN sirve para codificar la información genética de los organismos, a través de regular el autoensamble de biomoléculas como las proteínas, que llevan a cabo la mayoría de las funciones vitales.

Las proteínas son cadenas de aminoácidos, los cuales se pliegan en una conformación particular. Se sabe que la función que realiza la proteína está altamente relacionada con la estructura tridimensional, y a su vez, la estructura tridimensional de una proteína depende principalmente de la secuencia de aminoácidos que la componen.

El proceso de secuenciar las proteínas es más rápido que el de determinar su estructura tridimensional. Se conocen más de 11 millones de secuencias (uni, 2010b; Wu *et al.*, 2006), pero sólo 67,322 estructuras (rcs, 2010).

La estructura tridimensional de las proteínas se obtiene a través de métodos experimentales, de los cuales, los más utilizados y ampliamente aceptados son el método de cristalografía por rayos X (Gu y Bourne, 2009) y el método de resonancia magnética nuclear (Gu y Bourne, 2009).

El proceso experimental para determinar la estructura tridimensional de las proteínas es complejo y laborioso, puede tomar desde semanas hasta años generar un buen modelo.

De ahí surge la necesidad de crear métodos computacionales que predigan la estructura tridimensional a partir de la secuencia de aminoácidos. A esto se le conoce como el problema de **predicción de estructura**.

La solución al problema de predicción de estructura servirá para detectar enfermedades. Por ejemplo, si se presenta una mutación en un gen, esto puede provocar cambios en la secuencia de la proteína y modificar su función. Utilizando los métodos computacionales se puede determinar si la mutación ocasiona un cambio conformacional importante, que impacta su función y provoque una enfermedad, como es el caso de la anemia de células falciformes (ver Sección II.1.1).

Otro problema importante relacionado es el de **diseño de proteínas**. Éste es un problema inverso al de la predicción de estructura, y consiste en determinar la secuencia de aminoácidos que generan un plegamiento deseado.

En el diseño de proteínas, se propone una estructura deseada, y se debe determinar la secuencia que genera tal estructura. Una manera de tratar de resolver el problema es proponer secuencias en forma iterativa y determinar la estructura de cada secuencia, hasta conseguir la estructura deseada.

Por lo tanto se puede observar una clara relación entre ambos problemas. El problema conocido como *empaquetamiento de la cadena lateral de proteínas* (PSCPP) se aplica tanto en predicción de estructura como en el diseño de proteínas. La idea principal es utilizar las estructuras conocidas para predecir parte de la estructura de una nueva secuencia, y posteriormente pasar por una fase de refinamiento utilizando entre otras cosas, métodos que resuelven el PSCPP.

Muchas proteínas realizan sus funciones a través de átomos que quedan expuestos en la superficie de la proteína y que son parte de la cadena lateral, por lo que predecir el empaquetamiento de la cadena lateral es crucial para inferir la función de una proteína a partir de la estructura. A este problema se le conoce como **predicción de función**.

También existe lo que se conoce como **rediseño de proteínas**, esto consiste en que se tiene una proteína, con secuencia, estructura y función conocidas, las cuales se desean modificar para: aumentar su termoestabilidad[1], alterar la especificidad de enlace, aumentar su actividad enzimática, alterar la especificidad del sustrato, etc.

Las proteínas toman una conformación llamada nativa bajo condiciones normales; sin embargo, si la temperatura, presión, pH u otras variables del ambiente cambian, la proteína puede cambiar su conformación, y con ello su función.

Para los problemas mencionados (predicción de estructura, diseño de proteínas,

[1]La termoestabilidad está relacionada al intervalo de temperatura en el que puede existir una proteína antes de perder su confomación nativa.

predicción de función y rediseño de proteínas) se necesitan métodos que resuelvan el PSCPP.

El PSCPP se mapea a un problema de optimización combinatoria, y Akutsu (1997) demostró que este problema pertenece a la clase NP-difícil, por lo que se han desarrollado varias de heurísticas que aproximan el PSCPP.

Los métodos actuales más destacados son el SCWRL4 (Krivov *et al.*, 2009) y el OPUS-Rota (Lu *et al.*, 2008b). Actualmente se desconoce cuál es mejor y una comparación entre ambos aportaría información sobre cuál método utilizar en casos específicos. Un aspecto fundamemtal para la realización de esta comparación es la generación de casos de prueba. El trabajo aquí propuesto pretende llenar estos huecos.

A continuación se presentan los objetivos de la presente investigación.

I.2 Objetivos de la investigación

I.2.1 Objetivo general

Determinar el desempeño relativo de los algoritmos conocidos como OPUS-Rota y SCWRL4 cuando estos son aplicados a casos heterogéneos del PSCPP.

I.2.2 Objetivos específicos

- Determinar si la clasificación SCOP es un determinante para la calidad de las soluciones de los métodos que aproximan el PSCPP.
- Determinar si la clasificación EC es un determinante para la calidad de las soluciones de los métodos que aproximan el PSCPP.
- Determinar si el tipo de aminoácido es un determinante para la calidad de las soluciones de los métodos que aproximan el PSCPP.
- Determinar qué función es mejor para medir la calidad de las soluciones de los métodos que aproximan el PSCPP.

I.3 Organización de la tesis

El presente trabajo se divide en las siguientes secciones:

En el Capítulo II se da una introducción al tema de las proteínas, su composición química, estructura y funciones, y al banco de datos de las proteínas, en el cual se almacenan las estructuras de las mismas.

En el Capítulo III se presenta el PSCPP, la discretización del problema, antecedentes y medidas para determinar la calidad de las soluciones de los métodos que aproximan el PSCPP.

En el Capítulo IV se proponen los casos de prueba y se presentan los experimentos realizados para comparar los mejores métodos existentes.

En el Capítulo V se describen las conclusiones a las que se llegó, así como propuestas para la continuación de este trabajo de investigación.

En el Apéndice A se muestra las características de un algoritmo simple basado en la técnica de recocido simulado que aproxima el PSCPP.

Capítulo II

MARCO TEÓRICO

II.1 Biología de las proteínas

El desarrollo de la tecnología ha influido en el avance de muchas áreas del conocimiento, y la biología no es la excepción. Además, para el caso de la biología, gracias a la tecnología se han generado grandes cantidades de datos, los cuales, con ayuda de las ciencias de la computación, se deben transformar en información útil que genere conocimientos para la biología. El área transdisciplinaria que busca este fin es la biocomputación.

La biocomputación genera modelos que ayudan a predecir el comportamiento de los sistemas biológicos. Estos modelos se pueden ser basar en datos y en propiedades fisico-químicas de las biomoléculas.

La biocomputación es muy amplia y se puede ser dividir en diversas áreas tal como se muestra en la Figura 1. La biocomputación se enfoca en el desarrollo de conocimientos, principalmente para la biología molecular. Las grandes áreas de la biocomputación son genómica, proteómica y biología de sistemas.

La genómica es uno de los dominios más importantes en biocomputación, debido a que el número de secuencias de ADN disponibles está creciendo exponencialmente (Larranaga *et al.*, 2006) y en consecuencia estos datos necesitan procesarse para obtener información útil.

Si los genes contienen la información, las proteínas son los trabajadores que transforman dicha información en vida. Las proteínas juegan un rol muy importante en el proceso de la vida, y su estructura tridimensional (3D) es una característica clave en su función. En el dominio de la proteómica, las aplicaciones principales de métodos computacionales son: predicción de estructuras, predicción de funciones y diseño de proteínas. Las proteínas son macromoléculas muy complejas con miles de átomos y enlaces. Por lo tanto, el número de posibles estructuras es muy grande. Esto hace la predicción de estructuras de proteínas un problema combinatorio muy complicado donde se requiere el uso de técnicas de optimización.

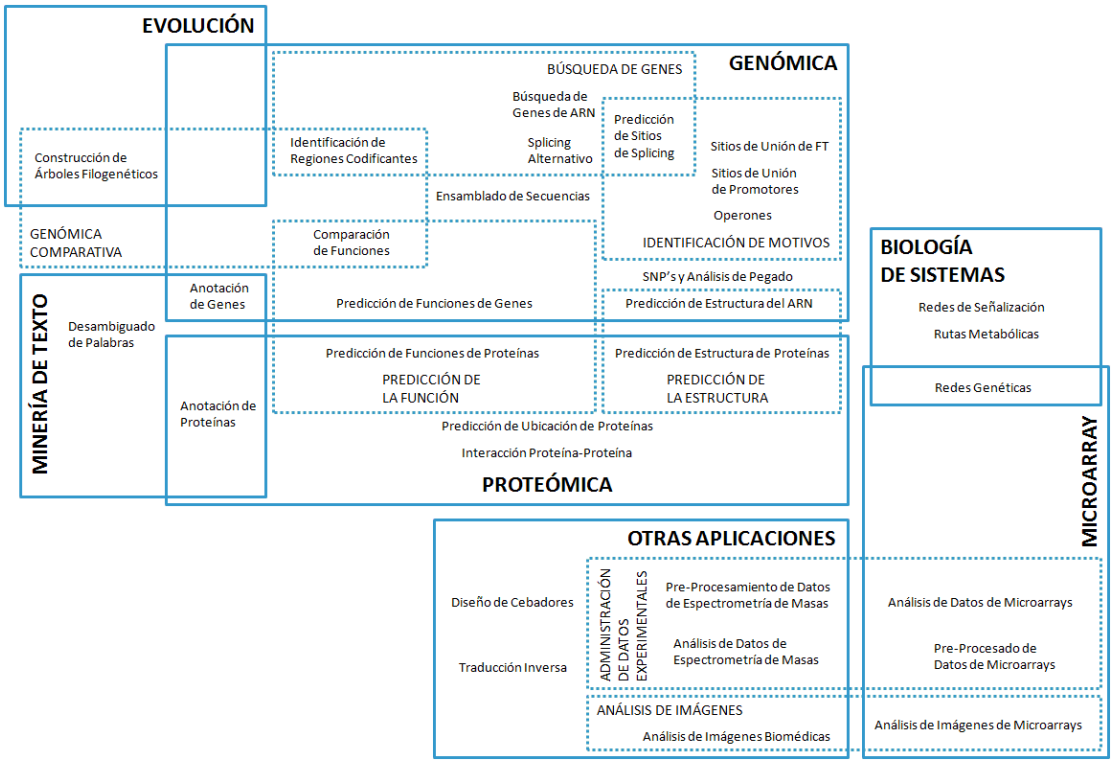


Figura 1. Áreas de biocomputación (Larranaga *et al.*, 2006).

Tabla I. Los 20 aminoácidos estándar y sus códigos.

Aminoácido	Código de tres letras	Código de una letra
Alanina	ALA	A
Cisteína	CYS	C
Ácido aspártico	ASP	D
Ácido glutámico	GLU	E
Fenilalanina	PHE	F
Glicina	GLY	G
Histidina	HIS	H
Isoleucina	ILE	I
Lisina	LYS	K
Leucina	LEU	L
Metionina	MET	M
Asparagina	ASN	N
Prolina	PRO	P
Glutamina	GLN	Q
Arginina	ARG	R
Serina	SER	S
Treonina	THR	T
Valina	VAL	V
Triptófano	TRP	W
Tirosina	TYR	Y

A continuación se detallan un poco más la composición, estructura y funciones de las proteínas, ya que son el objeto principal de estudio para esta investigación.

II.1.1 Proteínas

Las proteínas son macromoléculas constituidas por cadenas de aminoácidos unidos por enlaces peptídicos (Gu y Bourne, 2009), por lo que se les llama cadenas polipeptídicas. En la Tabla I se muestran los nombres y códigos de los 20 aminoácidos estándar que componen las proteínas, mientras que en la Figura 2 se muestra la estructura química de los mismos junto con el código de tres letras.

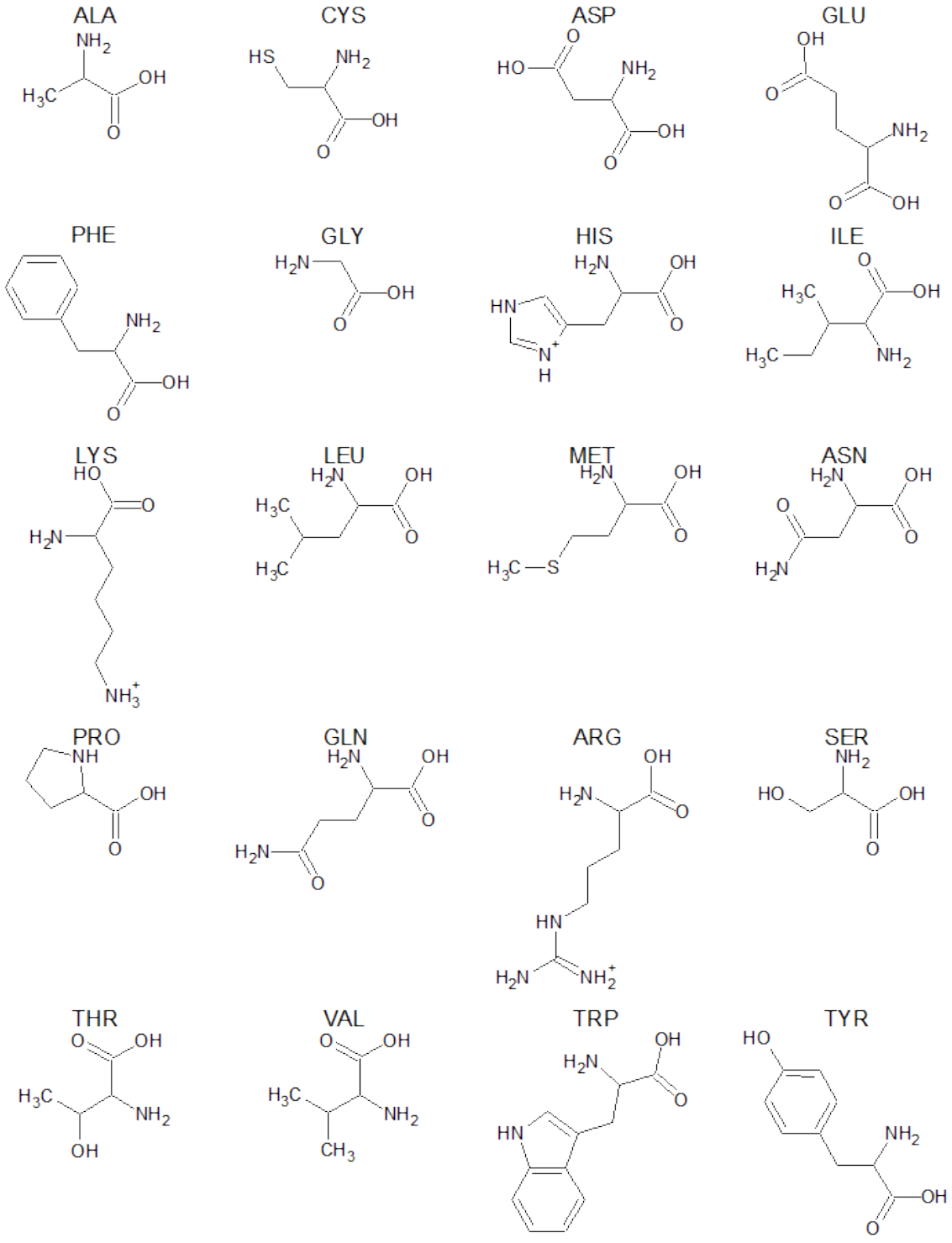


Figura 2. Aminoácidos que componen a las proteínas.

	U	C	A	G	
U	UUU PHE	UCU SER	UAU TYR	UGU CYS	U
	UUC PHE	UCC SER	UAC TYR	UGC CYS	C
	UUA LEU	UCA SER	UAA Parada	UGA Parada	A
	UUG LEU	UCG SER	UAG Parada	UGG TRP	G
C	CUU LEU	CCU PRO	CAU HIS	CGU ARG	U
	CUC LEU	CCC PRO	CAC HIS	CGC ARG	C
	CUA LEU	CCA PRO	CAA GLN	CGA ARG	A
	CUG LEU	CCG PRO	CAG GLN	CGG ARG	G
A	AUU ILE	ACU THR	AAU ASN	AGU SER	U
	AUC ILE	ACC THR	AAC ASN	AGC SER	C
	AUA ILE	ACA THR	AAA LYS	AGA ARG	A
	AUG MET	ACG THR	AAG LYS	AGG ARG	G
G	GUU VAL	GCU ALA	GAU ASP	GGU GLY	U
	GUC VAL	GCC ALA	GAC ASP	GGC GLY	C
	GUA VAL	GCA ALA	GAA GLU	GGA GLY	A
	GUG VAL	GCG ALA	GAG GLU	GGG GLY	G

Figura 3. Código Genético.

Las proteínas se forman a partir de la información contenida en el ADN de los organismos a través del proceso de **transcripción-traducción**. Este proceso se lleva a cabo de diferente manera según el organismo en el que se realice; sin embargo, en general, todos los organismos comparten el mismo **código genético**. En el proceso de transcripción-traducción, de manera simplificada, se traduce un gen a una cadena de ARN mensajero, el cual posteriormente, se lee en grupos de tres ácidos ribonucleicos llamados codones para generar un aminoácido específico que forma parte de la proteína que se está produciendo. La relación entre codones y aminoácidos, *i.e.*, código genético, se muestra en la Figura 3. Se sabe que las bases de los ácidos ribonucleicos son: uracilo (U), citosina (C), adenina (A) y guanina (G), con los cuales, agrupados de tres en tres, se pueden generar $4^3 = 64$ diferentes aminoácidos; sin embargo, el código está degenerado y produce sólo 20 aminoácidos.

Aminoácidos

Los 20 aminoácidos que componen a las proteínas tienen una estructura base común, pero difieren en sus cadenas laterales tal y como se indica en la Figura 4. Las cadenas laterales son las que hacen diferentes a los aminoácidos y le proveen sus características químicas.

En la Figura 4 se puede observar la estructura general de un aminoácido. La cadena

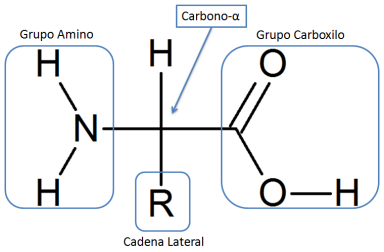


Figura 4. Estrucutra general de un aminoácido

lateral no se especifica ya que varía de un tipo de aminoácido a otro. La columna vertebral de un aminoácido se compone de un grupo carboxilo (-COOH), de un grupo amino ($-NH_2$) y un átomo de carbono central denominado carbono-α (C^{α}). La composición de la columna vertebral se mantiene en los 20 aminoácidos estándar.

Es difícil hacer una clasificación única de los aminoácidos, por lo que se pueden hacer distintas clasificaciones de acuerdo a diversos criterios. Por ejemplo, los aminoácidos se pueden clasificar según sus características químicas como se muestra en la Figura 5. Debido a que los aminoácidos tirosina (Y) y glicina (G) presentan propiedades químicas particulares, estos no se encuentran en la clasificación aquí mostrada.

Estas propiedades químicas son de gran importancia en el estudio de la predicción de estructuras, ya que hay teorías que indican que la propiedad química que tenga un aminoácido genera cierta afinidad con su entorno, es decir, por estar en el centro o en la superficie de la estructura tridimensional según sea el medio que los rodea.

Átomos de las Proteínas

Los aminoácidos se componen principalmente de carbono (C), nitrógeno (N), oxígeno (O), azufre (S) e hidrógeno (H). Los aminoácidos difieren entre ellos en su cadena lateral, por lo que es necesario saber cuáles y cuántos átomos tiene cada aminoácido. Los átomos de la cadena lateral se etiquetan de acuerdo a la cercanía que tienen respecto

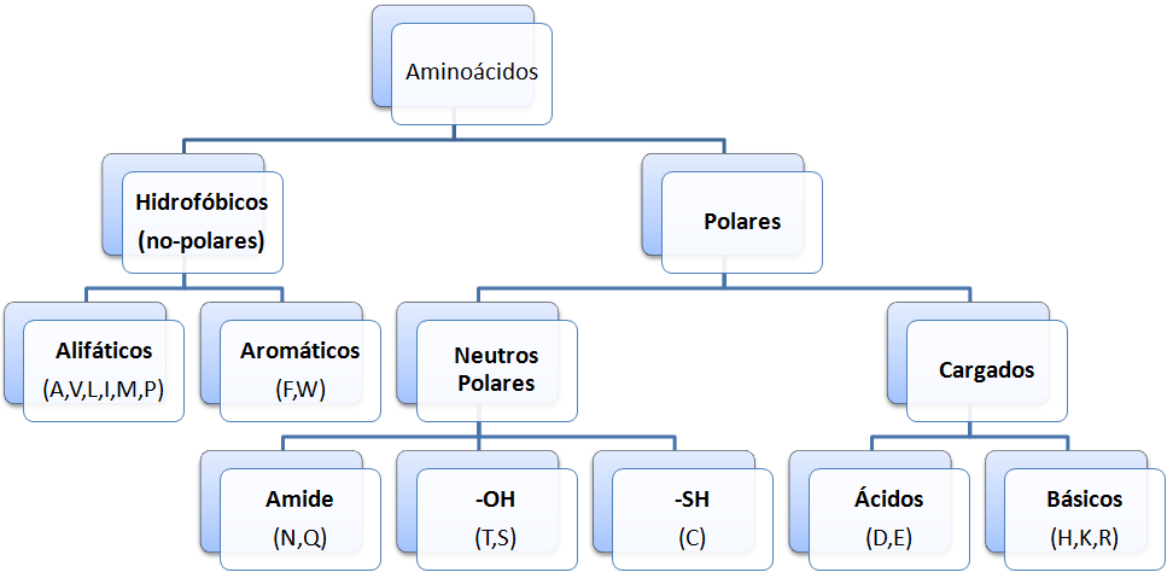


Figura 5. Clasificación de los aminoácidos según sus propiedades químicas.

al C^{α}, y se etiquetan con las letras griegas β, γ, δ, ϵ, η y ι, además, si hay dos átomos a la misma distancia, se etiquetan con números arábigos.

La lista de átomos pesados (C, N, O y S) por aminoácido se presenta en la Tabla II. Como se puede observar en dicha tabla, todos los aminoácidos tienen cuatro átomos en la columna vertebral; sin embargo, difieren en la cantidad de átomos de la cadena lateral. El aminoácido con menos átomos en la cadena lateral es la glicina, con cero átomos, mientras que el aminoácido con más átomos es el triptófano con diez átomos pesados.

Niveles Estructurales de las Proteínas

Para estudiar a las proteínas se tienen diferentes niveles de estructuras. Van desde las estructuras primarias hasta las quinarias, agregando complejidad en cada nivel.

La estructura más básica es la **estructura primaria**, que se define por el conjunto de enlaces covalentes de la proteína; por simplicidad, la estructura primaria se

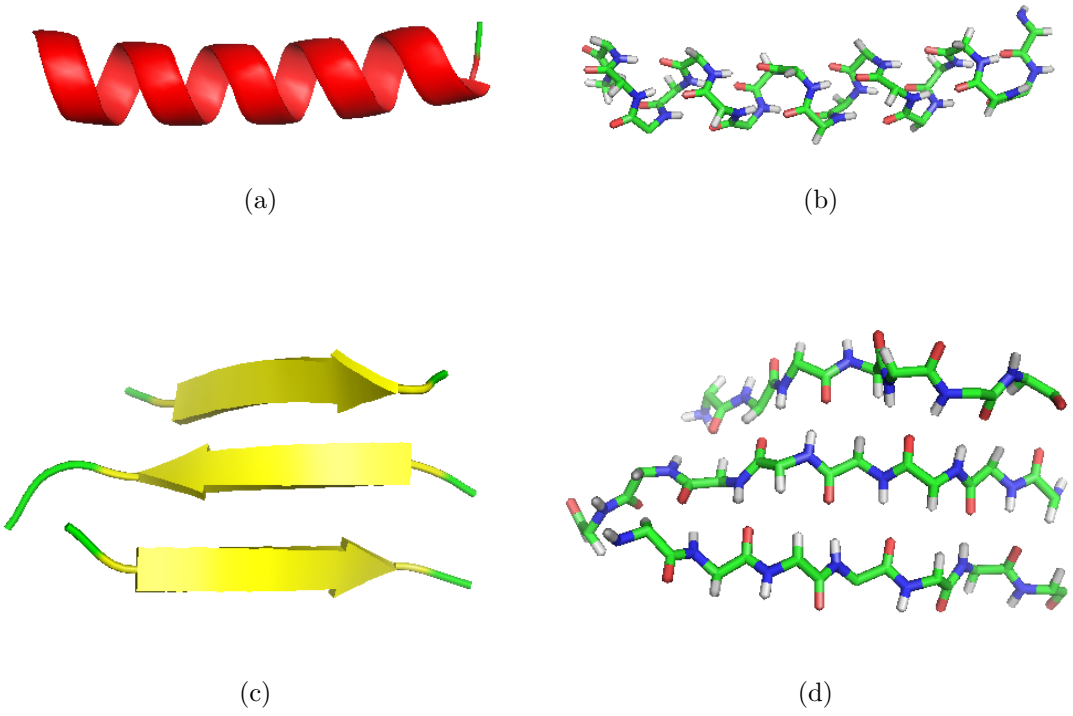
Tabla II. Los áminoácidos y sus átomos.

Aminoácido	Átomos Columna Vertebral	Átomos Cadena Lateral
ALA	N, O, C, C^{α}	C^{β}
ARG	N, O, C, C^{α}	C^{β}, C^{γ}, C^{δ}, N^{ϵ}, C^{η}, N_1^{ι}, N_2^{ι}
ASN	N, O, C, C^{α}	C^{β}, C^{γ}, O_1^{δ}, N_2^{δ}
ASP	N, O, C, C^{α}	C^{β}, C^{γ}, O_1^{δ}, O_2^{δ}
CYS	N, O, C, C^{α}	C^{β}, S^{γ}
GLN	N, O, C, C^{α}	C^{β}, C^{γ}, C^{δ}, O_1^{ϵ}, N_2^{ϵ}
GLU	N, O, C, C^{α}	C^{β}, C^{γ}, C^{δ}, O_1^{ϵ}, O_2^{ϵ}
GLY	N, O, C, C^{α}	—
HIS	N, O, C, C^{α}	C^{β}, C^{γ}, N_1^{δ}, C_2^{δ}, C_1^{ϵ}, N_2^{ϵ}
ILE	N, O, C, C^{α}	C^{β}, C_1^{γ}, C_2^{γ}, C_1^{δ}
LEU	N, O, C, C^{α}	C^{β}, C^{γ}, C_1^{δ}, C_2^{δ}
LYS	N, O, C, C^{α}	C^{β}, C^{γ}, C^{δ}, C^{ϵ}, N^{η}
MET	N, O, C, C^{α}	C^{β}, C^{γ}, S^{δ}, C^{ϵ}
PHE	N, O, C, C^{α}	C^{β}, C^{γ}, C_1^{δ}, C_2^{δ}, C_1^{ϵ}, C_2^{ϵ}, C^{η}
PRO	N, O, C, C^{α}	C^{β}, C^{γ}, C^{δ}
SER	N, O, C, C^{α}	C^{β}, O^{γ}
THR	N, O, C, C^{α}	C^{β}, O_1^{γ}, C_2^{γ}
TRP	N, O, C, C^{α}	C^{β}, C^{γ}, C_1^{δ}, C_2^{δ}, N_1^{ϵ}, C_2^{ϵ}, C_3^{ϵ}, C_2^{η}, C_3^{η}, C_2^{ι}
TYR	N, O, C, C^{α}	C^{β}, C^{γ}, C_1^{δ}, C_2^{δ}, C_1^{ϵ}, C_2^{ϵ}, C^{η}, O^{ι}
VAL	N, O, C, C^{α}	C^{β}, C_1^{γ}, C_2^{γ}

puede definir por la cantidad y el órden de aminoácidos que componen a una cadena polipeptídica, es decir, la secuencia de aminoácidos. Conocer la estructura primaria de las proteínas es importante para conocer la función de las mismas, así como para el estudio de enfermedades genéticas. Una proteína mal formada, con una estructura primaria diferente a la normal, puede provocar que el funcionamiento de la proteína no sea el adecuado o que su función no se realice y ocasionar así trastornos en los organismos que la generan (Gu y Bourne, 2009). Un ejemplo de esto es la enfermedad conocida como anemia de células falciformes, que consiste en una simple mutación (E6V[1]) en las cadenas B de la misma, la cual produce un cambio en la conformación global de la hemoglobina. Esta enfermedad provoca eventualmente daños en los órganos, acortando la esperanza de vida, 42 y 48 años para hombres y mujeres, respectivamente (Platt *et al.*, 1994).

La **estructura secundaria** hace referencia a las estructuras *hélices* α y *hojas* β que

[1]La hemoglobina consiste de cuatro cadenas, dos cadenas A y dos cadenas B. En la posición 6 de las cadenas B de una hemoglobina normal está un ácido glutámico (E), pero la mutación E6V significa que este residuo cambia por una valina (V), generando un cambio en la conformación de la hemoglobina.



(a) (b)

(c) (d)

Figura 6. Estructuras secundarias de las proteínas. (a) y (b) Diagrama de los residuos 40-59 de la proteína 1A3C que tienen una conformación de hélice, usando el modelo "cartoon" y "sticks" respectivamente con ayuda del programa "PyMol". (c) y (d) Diagrama de los residuos 3-9, 160-166 y 173-179 de la proteína 1A3C que tienen una conformación de hoja, usando el modelo "cartoon" y "sticks" respectivamente con ayuda del programa "PyMol".

definen algunas zonas de la columna vertebral de la proteína. En la Figura 6 se muestra un diagrama con la ejemplificación de las estructuras secundarias de una proteína.

La **estructura terciaria** se define por la estructura tridimensional global de la cadena polipeptídica. Este nivel estructural es el de interés para el presente proyecto de investigación. La manera en la que se pliegan las proteínas varía según el medio en el que éstas se encuentren; la conformación de mayor importancia es la que toma en su estado natural, es decir, en solución acuosa, donde se encuentra en el organismo y donde realiza sus funciones.

Los niveles estructurales más complejos, las estructuras cuaternarias y quinarias,

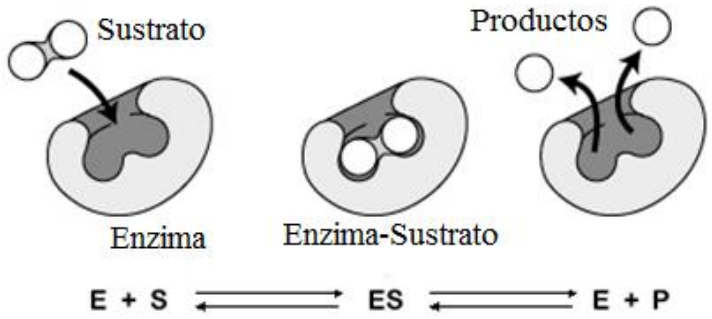


Figura 7. Ejemplo de proteína con función enzimática (Leja, 2010).

incorporan información de la interacción de varias cadenas polipeptídicas, ya que en ocasiones, una función la realizan más de una cadena a la vez.

Funciones de las Proteínas

Las proteínas intervienen en las funciones vitales de los organismos vivos. Las funciones se pueden categorizar en (Jensen *et al.*, 2002): defensa, reserva, hormonal, transporte, movimiento, reguladora, estructural, enzimática, traducción y reconocimiento de señales.

Una proteína puede tener más de una función, y se cree que las funciones están íntimamente relacionadas con la estructura terciaria de la misma. En la Figura 7 se muestra un diagrama de una proteína con función enzimática, donde su función la lleva a cabo mediante el uso de cavidades generadas por un plegamiento específico. La función de las enzimas es la de romper o crear enlaces químicos. Si hablamos de una enzima cuya función es romper enlaces, la reacción procede del estado "E+S" al estado "E+P". En esta reacción, el sustrato llega a la enzima, y ésta rompe los enlaces liberando dos o más productos.

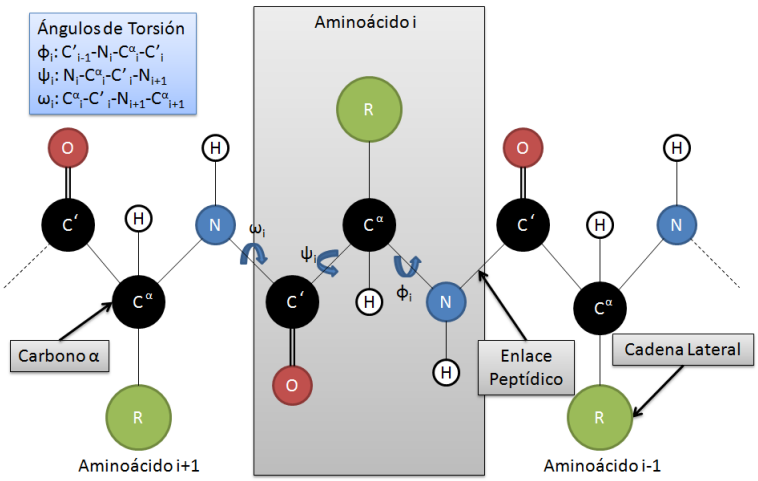


Figura 8. Ángulos de Torsión de la Cadena Principal.

II.1.2 Ángulos de torsión

Ángulos de Torsión de la Columna Vertebral

Como se dijo anteriormente, una proteína es una cadena de aminoácidos, los cuales, al unirse unos con otros, liberan una molécula de agua, por lo que queda solamente un átomo de nitrógeno, uno de oxígeno, y dos de carbono en la cadena principal. El átomo de carbono unido a la cadena lateral se denomina carbono-α (C^{α}), mientras que el otro se le puede etiquetar como el carbono C' o simplemente C. En la Figura 8 se muestra, de manera simplificada, el i-ésimo aminoácido de una proteína cualquiera, junto con los aminoácidos $i-1$ e $i+1$.

La proteína es un objeto geométrico, el cual se representa con las coordenadas de los átomos que la componen; si la proteína se somete a traslación o rotación, la estructura global no sufre cambios pero las coordenadas sí. Por lo tanto, se puede representar también por los ángulos de torsión que se forman con los átomos, y estos ángulos no se ven afectados por rotación ni traslación.

Un ángulo de torsión se forma por cuatro puntos en el espacio (i, j, k y $l \in \Re^3$) que

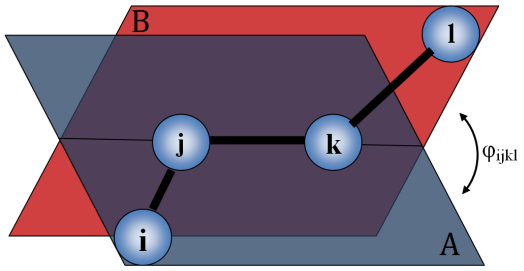


Figura 9. Ángulo de torsión de los planos formados por los puntos i, j, k y l. Los puntos i, j, k definen el plano A y los puntos j, k, l el plano B. φ_{ijkl} es el ángulo entre la normal al plano A n_A y la normal al plano B n_B.

representan a las coordenadas x, y, z de los cuatro átomos (ver Figura 9). Los puntos están relacionados por pares $((i,j), (j,k), (k,l))$. Los puntos i, j y k definen al plano A, y los puntos j, k y l definen al plano B. Una manera de calcular el ángulo de torsión φ_{ijkl} entre los planos A y B es mediante los vectores normales unitarios n_A y n_B a los planos A y B, respectivamente (Ecuación 1).

$$\varphi_{ijkl} = \cos^{-1}((n_A) \cdot (n_B)) \tag{1}$$

Sean $b_1 = j - i$, $b_2 = k - j$ y $b_3 = l - k$, entonces el ángulo de torsión (φ_{ijkl}) de los átomos i, j, k y l se define también por la Ecuación 2.

$$\varphi_{ijkl} = \tan_2^{-1}(|b_2|b_1 \cdot [b_2 \times b_3], [b_1 \times b_2] \cdot [b_2 \times b_3]) \tag{2}$$

La función $\tan_2^{-1}(\cdot,\cdot)$ se define como:

$$\tan_2^{-1}(y,x) = \begin{cases} \tan^{-1}|\frac{y}{x}| \cdot sgn(y) & x > 0, y \neq 0 \\ \frac{\pi}{2} \cdot sgn(y) & x = 0, y \neq 0 \\ (\pi - \tan^{-1}|\frac{y}{x}|) \cdot sgn(y) & x < 0, y \neq 0 \\ 0 & x > 0, y = 0 \\ \text{indefinido} & x = 0, y = 0 \\ \pi & x < 0, y = 0 \end{cases} \quad (3)$$

En la Figura 8 se muestran los ángulos de la columna vertebral, indicados en el enlace donde se genera el ángulo de torsión. Los ángulos de torsión de la columna vertebral son tres por aminoácido (con excepción de los aminoácidos inicial y final), los cuales se listan a continuación para el i-ésimo residuo junto con los átomos que forman el ángulo indicado.

- ϕ_i: C'$_{i-1}$ - N$_i$ - C$_i^{\alpha}$ - C'$_i$
- ψ_i: N$_i$ - C$_i^{\alpha}$ - C'$_i$ - N$_{i+1}$
- ω_i: C$_i^{\alpha}$ - C'$_i$ - N$_{i+1}$ - C$_{i+1}^{\alpha}$

Además de estos ángulos, las cadenas laterales también forman ángulos de torsión, los cuales son los que definen el empaquetamiento. A continuación se describen los ángulos de torsión de la cadena lateral.

Ángulos de Torsión de la Cadena Lateral

En esta sección se describen los ángulos de torsión de la cadena lateral. Para ello, se utiliza el ejemplo del ácido aspártico (ASP). El ASP, debido a su longitud, tiene sólo dos ángulos de torsión en su cadena lateral, que son los ángulos χ_1 y χ_2. La relación de la cantidad de ángulos por residuo se muestra en la Tabla III.

Tabla III. Cantidad de ángulos de torsión de la cadena lateral por aminoácido.

Aminoácido	χ_1	χ_2	χ_3	χ_4	χ_5
ALA	×	×	×	×	×
ARG	✓	✓	✓	✓	✓
ASN	✓	✓	×	×	×
ASP	✓	✓	×	×	×
CYS	✓	×	×	×	×
GLN	✓	✓	✓	×	×
GLU	✓	✓	✓	×	×
GLY	×	×	×	×	×
HIS	✓	✓	×	×	×
ILE	✓	✓	×	×	×
LEU	✓	✓	×	×	×
LYS	✓	✓	✓	✓	×
MET	✓	✓	✓	×	×
PHE	✓	✓	×	×	×
PRO	✓	✓	×	×	×
SER	✓	×	×	×	×
THR	✓	×	×	×	×
TRP	✓	✓	×	×	×
TYR	✓	✓	×	×	×
VAL	✓	×	×	×	×

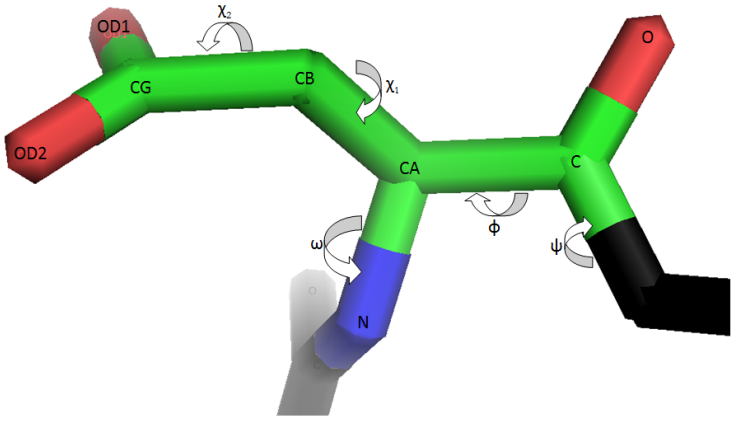


Figura 10. Ángulos de torsión del ASP. El diagrama sigue el estándar de colores de los átomos (azul = nitrógeno, rojo = oxígeno).

En la Figura 10 se muestra el ejemplo del ácido aspártico, en el cual, los ángulos χ_1 y χ_2 se calculan con los átomos N - C^{α} - C^{β} - C^{γ} y C^{α} - C^{β} - C^{γ} - O_1^{δ}, respectivamente.

La relación general para cada uno de los ángulos de torsión de la cadena lateral es la siguiente:

- χ_1: N - C^{α} - C^{β} - X_1^{γ}
- χ_2: C^{α} - C^{β} - X_1^{γ} - X_1^{δ}

- χ_3: C^{β} - X_1^{γ} - X_1^{δ} - X_1^{ϵ}
- χ_4: X_1^{γ} - X_1^{δ} - X_1^{ϵ} - X_1^{η}
- χ_5: X_1^{δ} - X_1^{ϵ} - X_1^{η} - X_1^{ι}

donde N y C^{α} son átomos de la cadena principal, y los demás son átomos de la cadena lateral. Debido a que los aminoácidos difieren entre ellos en la cadena lateral, los ángulos de torsión de la cadena lateral se calculan según el aminoácido del cual se esté tratando. Como se puede ver en la Tabla II, todos los aminoácidos inician su cadena lateral con el C^{β}, por lo que en la definición de los ángulos χ_1, χ_2 y χ_3 se utiliza explícitamente al C^{β} para indicar al átomo que le sigue al C^{α}. Por el contrario, después del C^{β} pueden ir diferentes tipos de átomos (N, O o C), por lo que en la definición de los ángulos de torsión se indican como X_1^{γ} para indicar que es el primer átomo que le sigue al C^{β}. De la misma manera se utiliza la nomenclatura de X_1^{δ}, X_1^{ϵ}, X_1^{η} y X_1^{ι} para indicar que son los átomos subsecuentes.

Se puede ver en el ejemplo del ácido aspártico, que de la definición general de los ángulos de torsión, se reemplazó a X_1^{γ} por el primer átomo γ del aminoácido, que es el C^{γ}, y de la misma manera, se reemplazó a X_1^{δ} como el primer átomo que le sigue a X_1^{γ}, que en este caso es O_1^{δ}.

Para el presente problema se necesita, además, obtener las coordenadas de átomos a partir de los ángulos de torsión. El prodecimiento para realizar esto se describe en la Sección A.1.5.

II.1.3 PDB (*Protein Data Bank*)

Como una iniciativa para almacenar las estructuras de las proteínas, surgió el PDB (Bernstein *et al.*, 1978). En esta sección se explica el formato de los archivos que

almacenan las estructuras en este banco de datos.

El PDB es un sistema que sirve para almacenar y mantener las estructuras de biomoléculas como proteínas, ADN y ARN. Las estructuras se obtienen a través de métodos experimentales y se reportan en archivos de texto que tienen un formato específico (el formato PDB).

A través de los años el formato PDB ha sufrido modificaciones, esto debido a que en un inicio no se consideraron aspectos que surgieron con el tiempo, como incluir los datos de los métodos experimentales, la estructura secundaria, clasificación, etc.

Los archivos PDB tienen errores y por ello se tiene que ser cuidadoso con su uso. Se han registrado errores graves en estructuras reportadas (Gu y Bourne, 2009), y se está trabajando en depurar el PDB; sin embargo, la tarea de verificación puede llegar a ser igual de lenta que el mismo proceso de determinar la estructura terciaria nuevamente.

Para reportar una estructura en el PDB, los autores le asignan un identificador de cuatro caracteres, donde el primero es un dígito (1-9) y los demás caracteres son alfanuméricos. Este identificador (PDB ID) es único, es decir, dos estructuras diferentes no pueden tener el mismo PDB ID. Sin embargo, se da el caso que para una misma proteína se hayan generado diferentes estructuras experimentalmente, generadas tal vez por diferentes autores, o bajo diferentes condiciones (una enzima con ligando y sin ligando); entonces una misma proteína puede tener más de una estructura reportada, y tendrá entonces un PDB ID por cada una de ellas.

Otro aspecto interesante es que las proteínas se pueden fraccionar para su estudio, y un archivo PDB puede ser sólo una sección de la proteína completa, o por el contrario, un archivo PDB puede contar con más de una proteína o más de una cadena. Se le llama cadena a una secuencia ininterrumpida de aminoácidos en donde sólo hay una terminal N y una terminal C. Un archivo PDB puede contar con más de una cadena.

Es común que con los métodos experimentales no se pueda determinar la posición de algunos residuos, esto se debe a varios factores como flexibilidad de la proteína, o problemas con los experimentos. Por lo tanto, se pueden encontrar archivos PDB que tengan residuos faltantes, los cuales se pueden identificar de varias maneras. Una forma de identificar los residuos faltantes es a través del campo `REMARK 456` donde el autor anota los residuos faltantes de la estructura reportada, aunque este campo no es obligatorio para los autores. Otra forma es comparando el identificador de secuencia de los residuos; sin embargo, esto tampoco es tan confiable, ya que el identificador de secuencia de los residuos no es necesariamente contiguo. La manera más confiable para determinar si hay residuos faltantes, es calcular la distancia entre el átomo C del i-ésimo residuo y el átomo N del siguiente residuo, y compararlo con los valores promedio del enlace peptídico. Si esta distancia está lejos del promedio, entonces existen residuos faltantes.

Los archivos PDB cuentan con diversa información de la proteína reportada así como datos que se usaron en los experimentos para obtener la estructura. En especial, para la presente investigación los campos de mayor importancia son los siguientes.

Secuencia de aminoácidos (`SEQRES`)

A pesar de que existe otra base de datos para las secuencias de aminoácidos de las proteínas, en los archivos PDB se reporta la secuencia de aminoácidos para cada una de las cadenas del archivo. A través de la secuencia es como se puede determinar si existe más de un PDB ID para la misma proteína.

En la Tabla IV se muestra el formato que siguen los registros de las secuencias de aminoácidos en un archivo PDB. En la Figura 11 se muestra un extracto del archivo PDB cuyo PDB ID es 1A6F, la cual muestra registros de las secuencias de aminoácidos,

Tabla IV. Formato PDB para el registro de la secuecia de aminoácidos.

Columnas	Contenido
1-6	SEQRES
9-10	Número serial del registro SEQRES para la cadena actual
12	Identificador de la cadena
14-17	Número de residuos en esta cadena
20-22	Nombre del residuo
24-26	Nombre del residuo
⋮	
68-70	Nombre del residuo

```
1234567890123456789012345678901234567890123456789012345678901234567890
SEQRES   1 A  119  MET ALA HIS LEU LYS LYS ARG ASN ARG LEU LYS LYS ASN
SEQRES   2 A  119  GLU ASP PHE GLN LYS VAL PHE LYS HIS GLY THR SER VAL
SEQRES   3 A  119  ALA ASN ARG GLN PHE VAL LEU TYR THR LEU ASP GLN PRO
SEQRES   4 A  119  GLU ASN ASP GLU LEU ARG VAL GLY LEU SER VAL SER LYS
SEQRES   5 A  119  LYS ILE GLY ASN ALA VAL MET ARG ASN ARG ILE LYS ARG
SEQRES   6 A  119  LEU ILE ARG GLN ALA PHE LEU GLU GLU LYS GLU ARG LEU
SEQRES   7 A  119  LYS GLU LYS ASP TYR ILE ILE ILE ALA ARG LYS PRO ALA
SEQRES   8 A  119  SER GLN LEU THR TYR GLU GLU THR LYS LYS SER LEU GLN
SEQRES   9 A  119  HIS LEU PHE ARG LYS SER SER LEU TYR LYS LYS SER SER
SEQRES  10 A  119  SER LYS
```

Figura 11. Ejemplo de registros de secuencia de aminoácidos.

y se incluye una numeración para las columnas.

Coordenadas de Átomos (ATOM)

La información más importante de los archivos PDB son las coordenadas de los átomos. Las coordenadas de los átomos están dadas como coordendas ortogonales (x, y, z), y la magnitud está dada en Angstroms (Å)[2]. La información de las coordenadas de un átomo está dada en un registro (renglón) del archivo. En la Tabla V se muestra el formato que siguen los registros de los átomos en un archivo PDB. En la Figura 12 se muestra un extracto del archivo PDB cuyo PDB ID es 1A6F, la cual muestra registros

[2] 1 Å$= 1 \times 10^{-10}$ m

Tabla V. Formato PDB para el registro de las coordenadas de los átomos.

Columnas	Contenido
1-4	`ATOM`
7-11	Número serial del átomo
13-16	Nombre del átomo
17	Indicador de ubicación alternativa
18-20	Nombre del residuo
22	Identificador de la cadena
23-26	Número secuencial del residuo
27	Código para inserción de residuos
31-38	X
39-46	Y
47-54	Z
55-60	Ocupación
61-66	Factor de Temperatura
77-78	Símbolo del Elemento (Justificado a la derecha)
79-80	Carga en el átomo

de coordenadas de átomos, en el cual se incluye una numeración para las columnas.

En la mayoría de los archivos PDB se reportan sólo las coordenadas de los átomos pesados (C, O, N, S), aunque algunos reportan también las coordenadas de los átomos de hidrógeno. También hay estructuras que reportan únicamente las coordenadas del C_α, pero esto es cada vez menos común.

En las columnas 55-60 y 61-66 del registro `ATOM` se indica la ocupación y el factor de temperatura respectivamemte.

La **ocupación** de un átomo es regularmente 1.00 y significa que las cordenadas del átomo indicadas en ese registro (renglón) son únicas; sin embargo, se pueden tener varios registros (coordenadas) para un mismo átomo, y la ocupación estará distribuida entre todos los registros del mismo átomo, de tal manera que la suma de las ocupancias

```
123456789012345678901234567890123456789012345678901234567890123456
ATOM      1  N   ALA A   2      -4.360  57.812  -6.190  1.00 51.31
ATOM      2  CA  ALA A   2      -5.334  58.923  -6.353  1.00 51.89
ATOM      3  C   ALA A   2      -6.776  58.419  -6.577  1.00 52.43
ATOM      4  O   ALA A   2      -7.706  58.935  -5.963  1.00 53.24
ATOM      5  CB  ALA A   2      -4.891  59.859  -7.479  1.00 50.23
ATOM      6  N   HIS A   3      -6.962  57.400  -7.415  1.00 52.47
ATOM      7  CA  HIS A   3      -8.303  56.848  -7.667  1.00 51.84
ATOM      8  C   HIS A   3      -8.537  55.493  -6.982  1.00 51.33
ATOM      9  O   HIS A   3      -7.583  54.818  -6.608  1.00 51.33
ATOM     10  CB  HIS A   3      -8.542  56.737  -9.169  1.00 50.67
ATOM     11  CG  HIS A   3      -8.407  58.041  -9.876  1.00 51.74
ATOM     12  ND1 HIS A   3      -9.425  58.956 -10.005  1.00 52.15
ATOM     13  CD2 HIS A   3      -7.311  58.638 -10.400  1.00 52.59
ATOM     14  CE1 HIS A   3      -8.936  60.057 -10.573  1.00 52.68
ATOM     15  NE2 HIS A   3      -7.651  59.918 -10.835  1.00 54.78
```

Figura 12. Ejemplo de registros de coordenadas de átomos.

de todos los registros del mismo átomo sea uno. Si se tiene que seleccionar un conjunto de coordenadas para representar al átomo, se toma el que tenga mayor ocupación.

Este dato se obtiene experimentalmente y se relaciona con moleculas que pueden presentar consistentemente varias conformaciones. Por ejemplo, las porciones de la proteína que se adhieren a iones metálicos que provocan un cambio en la conformación para una porción de la proteína, así que se reportan varias conformaciones (coordenadas) y se indica el porcentaje de veces que ocurrió cada conformación. Algunos ejemplos de valores reportados para átomos con dos conformaciones son: 0.5 y 0.5, 0.6 y 0.4, 0.75 y 0.25.

El **factor de temperatura** es un indicador del modelo que representa el intervalo de la distribución de la densidad electrónica de un átomo obtenida experimentalmente. En una situación ideal, dónde un átomo esté en la misma posición para las diferentes moléculas del cristal, el intervalo de la distribución es pequeño y también el factor de temperatura. Sin embargo, en algunas situaciones, debido a la vibración de los átomos y algunas diferencias de conformación entre las diferentes moléculas del cristal,

la distribución de densidad de electrones es más amplia.

Un factor de temperatura menor a 10 representa un átomo cuya densidad de electrones fue puntual, y por el contrario, los registros cuyo factor de temperatura es mayor a 50 representan átomos con mucha movilidad, caso común para átomos de las cadenas laterales.

El proceso de mantenimiento del PDB es complejo. Hay varios pasos de validación para asegurar que tanto el formato como la información esté libre de errores. En el sitio web del PDB[3] se pueden descargar los archivos PDB. También cuenta con herramientas de búsqueda para filtrar utilizando diferentes criterios, como por ejemplo: PDB ID, número de cadenas, clasificación SCOP, clasificación CATH, longitud de las cadenas, clasificación EC, datos del experimento, tipo de método experimental utilizado, entre otros.

En el sitio web se pueden visualizar las proteínas, ya que cuenta con programas embebidos para visualización de estructuras como el Jmol; también cuenta con estadísticas sobre los archivos PDB, que ayudan a ver el crecimiento del número de estructuras por año, y los métodos experimentales más utilizados.

Para este trabajo se utilizó principalmente la búsqueda avanzada del sitio web para obtener los conjuntos de prueba filtrando por diferentes campos.

II.1.4 SCOP (*Structural Classification of Proteins*)

La clasificación SCOP es una clasificación jerárquica de estructuras de proteínas que se creó con el fin de agrupar las proteínas según sus características estructurales. La unidad de clasificación del SCOP es el dominio, que es un nivel más bajo que la cadena de una proteína, ya que una cadena puede estar constituída por uno o más dominios,

[3]http://www.pdb.org

pero a su vez, hay dominios que abarcan más de una cadena.

La clasificación SCOP no es la única clasificación de estructuras de proteínas, también se encuentra la clasificación CATH, pero difieren en el método de anotar las estructuras, ya que CATH es semi-automático y SCOP es manual, y utiliza la experiencia de profesionales para determinar la similitud entre un dominio y otro para su clasificación.

Ambos sistemas de clasificación pueden diferir al momento de clasificar una estructura, desde en la cantidad de dominios en las que se divide la proteína, hasta las clases que manejan cada una de estos sistemas. Se seleccionó la clasificación SCOP para este trabajo ya que utiliza un sistema de anotación manual y se consideró más conveniente, además de que su uso es más sencillo ya que almacena la información de la clasificación en archivos que se pueden analizar fácilmente (*parseable files*).

El proceso de anotación de estructuras en SCOP es más lento que la determinación de estructuras. En la versión 1.75 de la clasificación SCOP se tienen clasificadas 38222 estructuras (Brenner, 2009), mientras que en el PDB se tienen cerca de 70 mil estructuras.

Los niveles de clasificación de SCOP son:

1. Clase (*Class*)
2. Pliegue (*Fold*)
3. Superfamilia (*Superfamily*)
4. Familia (*Family*)

Por ejemplo, la estructura con PDB ID = 1fpo, tiene 6 dominios (d1fpoa1, d1fpob1, d1fpoc1, d1fpoa2, d1fpob2 y d1fpoc2), dónde los dóminios d1fpoa1, d1fpob1 y d1fpoc1

están en la clasificación a.2.3.1 y los dominios d1fpoa2, d1fpob2 y d1fpoc2 están en la clasificación a.23.1.1. Todos los dominios están en la clase a por lo que podemis decir que la proteína completa es de la clase a. A continuación se listan las clases de la clasificación SCOP.

- **clase a**, proteínas cuyas estructuras secundarias predominantes son hélices-α;
- **clase b**, proteínas cuyas estructuras secundarias predominantes son hojas-β;
- **clase c** proteínas con hélices-α y hojas-β (no separables); y
- **clase d**, proteínas con hélices-α y hojas-β (separables).
- **clase e**, proteínas multi-dominio.
- **clase f**, proteínas y péptidos de membrana y de la superficie de las células.
- **clase g**, proteínas pequeñas.
- **clase h**, proteínas en espiral.
- **clase i**, proteínas de baja resolución.
- **clase j**, peptidos.
- **clase k**, proteínas diseñadas.

Las clases más abundantes e importantes son las clases a, b, c y d, por lo que en nuestro estudio usaremos sólo estas clases e ignoraremos proteínas que no tengan clasificación SCOP o estén clasificadas bajo las clases e-k.

A continuación se explica la clasificación EC, que es otro tipo de clasificación de proteínas.

II.1.5 Clasificación EC (*Enzyme Comission*)

Para clasificar las proteínas se puede usar diferentes criterios según la característica que queramos estudiar. En la sección anterior se habla sobre un tipo de clasificación de proteínas de acuerdo con su estructura, aunque también se puede clasificar las proteínas de acuerdo con su función.

El número EC es la clasificación de proteínas de acuerdo con la función enzimática que realice. La clasificación EC es jerárquica, teniendo en el primer nivel 6 clases o grupos que a continuación se enlistan.

1. Oxidoreductasas
2. Transferasas
3. Hidrolasas
4. Liasas
5. Isomerasas
6. Ligasas

Por ejemplo, la proteína con PDB ID=1JVA es una endonucleasa con número EC 3.6.1.34. Dónde el número 3 representa el grupo de las hidrolasas y el resto es la subclase (Moss, 2010). Cabe señalar que este tipo de clasificación es sólo para proteínas con función enzimática conocida.

La clasificación EC se puede obtener en el archivo PDB de la estructura, a diferencia de la clasificación SCOP que es independiente y para relacionar un PDB con la clasificación SCOP se hace uso de los *parseable files*[4].

[4]http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html

II.1.6 Calidad de estructuras de proteínas

Las estructuras almacenadas en el PDB pueden tener errores, originados en alguno de los muchos pasos que se necesitan para determinar la estructura experimentalmente, ya sea por errores humanos, de interpretación de los resultados experimentales, o limitantes de la tecnología utilizada en el experimento. Por ello, se almacenan junto con las estructuras, parámetros que indiquen la calidad global del modelo. Entre los parámetros que indican la calidad de las estructuras están la resolución y el factor-R.

Resolución.

La resolución está relacionada con la longitud de onda utilizada para determinar la posición de los átomos en el método de cristalografía con rayos X. De manera general, se dice que un modelo tiene mejor resolución mientras menor sea el valor de ésta. Mientras menor sea la resolución del experimento, se podrá observar mayor nivel de detalle en las proteínas cristalizadas. Por ejemplo, el átomo más pequeño es el hidrógeno, cuyo radio[5] es de 1.2 Å, por lo que una resolución de 3 Å o mayor puede omitir el detalle de estos átomos, y con una resolución de 2 Å o menor, se puede apreciar mejor todos los átomos en el cristal, lo cual conlleva a un mejor modelo.

Factor R (R-Factor).

El factor R, al igual que la resolución, se utilizan para indicar la calidad del modelo. Este factor mide la similitud de los datos de difracción, obtenidos expermentalmente, y el modelo, el cual pasó por un proceso de refinamiento (Ecuación 4).

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{4}$$

[5]Radio de Van der Waals

En el proceso de refinamiento se utiliza la función del factor R como función objetivo, y se intenta minimizar y por esto, se dice que el factor-R esta sesgado. Para tener un parámetro libre de sesgo, se propone el factor R-free. El factor R-free se obtiene al optimizar el 90% (aproximadamente) de los datos en el proceso de refinamiento, y el 10% se utiliza para medir el factor R-free que indica que tan bien predice este 10% de los datos el nuevo modelo calculado.

Los valores recomendados para los parámetros factor-R y factor R-free son 0.2 o menos y 0.3 o menos, respectivamente (Gu y Bourne, 2009).

Capítulo III

DEFINICIÓN DEL PROBLEMA

El problema conocido como empaquetamiento de la cadena lateral en proteínas (PSCPP) se presenta en dos problemas de bioinformática principalmente. El primero, es el modelado homólogo, que consiste en la predicción de la estructura tridimensional de secuencias de proteínas a partir de estructuras de secuencias homólogas. El segundo problema es el rediseño de proteínas, en el cual se quiere modificar parte de una estructura conocida para que tenga diferentes propiedades, como aumentar su termoestabilidad, alterar su especificidad de enlace, mejorar su afinidad de enlace, aumentar su actividad enzimática, alterar la especificidad del sustrato, etc.

El PSCPP consiste en predecir la conformación de las cadenas laterales de una proteína dadas: la secuencia de aminoácidos y las coordenadas de los átomos de la columna vertebral de la misma. Lo que se requiere como salida para este problema son las coordenadas tridimensionales de cada uno de los átomos de las cadenas laterales; adicionalmente los átomos están sujetos a restricciones dadas por los enlaces entre átomos contiguos, pero estas restricciones no son suficientes para que el espacio de búsqueda sea discreto, pero Janin *et al.* (1978) han demostrado que las conformaciones de las cadenas laterales no siguen una distribución uniforme, sino que están agrupadas en ciertas regiones.

Se sabe que una proteína no tiene una única estructura en la naturaleza, aunque las conformaciones que pueda tomar en algún momento dado son estados de mínima energía libre. La energía libre de una proteína es difícil de calcular, debido a que intervienen numerosos factores y no se ha determinado una función que modele fielmente el plegamiento (Plaxco *et al.*, 2000; Gu y Bourne, 2009), por lo que se proponen distintas funciones de energía (Yanover *et al.*, 2008) que se utilizan para mapear el PSCPP a un problema de optimización.

Cabe recalcar que por eso mismo, a pesar de que existen métodos exactos que devuelven el óptimo global de las funciones propuestas, estos aún tienen un porcentaje apreciable de error en sus predicciones.

A continuación se describe a detalle cada uno de los aspectos de la transformación del PSCPP a un problema de optimización combinatoria.

III.1 Discretización del problema

III.1.1 Espacio de búsqueda

Como se dijo anteriormente, para el PSCPP se requiere encontrar las conformaciones de cada una de las cadenas laterales de los residuos de la proteína. Pero como se tienen restricciones de enlaces covalentes que limitan el espacio de búsqueda, es más fácil utilizar los ángulos de torsión de las cadenas laterales como las variables que se modifican para encontrar el óptimo de una función de energía establecida.

Algunos estudios han demostrado que estos ángulos no están distribuídos uniformemente en las estructuras conocidas hasta el momento, sino que hay regiones en las cuales se agrupan estos ángulos debido a que son las regiones con mínima energía (Gu y Bourne, 2009). De estos estudios se ha obtenido una de las técnicas de discretización de los ángulos de torsión de las cadenas laterales más utilizada, la cual se denomina bibliotecas de rotámeros. Existen otras técnicas basadas en una biblioteca de conformeros (Shetty *et al.*, 2003), y la discretización de los ángulos a través de pasos fijos, pero la técnica más popular es la biblioteca de rotámeros (Canutescu *et al.*, 2003; Peterson *et al.*, 2004; Liang y Grishin, 2002; Jain *et al.*, 2006; Wang *et al.*, 2005; Xiang y Honig, 2001; Krivov *et al.*, 2009).

A continuación se describe la técnica de discretización mediante la biblioteca de rotámeros.

Biblioteca de rotámeros

Una biblioteca de rotámeros se genera estadísticamente. Se utilizan proteínas con estructura conocida y de ellas se obtienen los diferentes ángulos de torsión de las cadenas laterales, agrupados por tipo de aminoácido. Las bibliotecas de rotámeros entonces son colecciones de rotámeros los cuales tienen asociados una probabilidad que está relacionada a la frecuencia que tiene el rotámero en el conjunto utilizado para su obtención.

A través de los años se han publicado diferentes bibliotecas de rotámeros debido a que cada día se tienen disponible más estructuras que se utilizan para generar mejores

bibliotecas y también se utilizan diferentes métodos para la obtención de las mismas.

Se ha demostrado (Lu *et al.*, 2008b) que el generar bibliotecas muy grandes no necesariamente genera mejores resultados, ya que al momento de utilizar la biblioteca para la predicción puede ocasionar que el método de búsqueda tarde más en encontrar las conformaciones que minimicen la energía libre. Sin embargo, las bibliotecas que se generaron utilizando muy pocas estructuras puede llevar a generar una biblioteca que no tenga ángulos representativos y reduzca la calidad de predicción.

Algunos autores opinan que la coformación de la cadena lateral tiene una relación directa con la conformación de la columna vertebral (Dunbrack y Karplus, 1993), por esto se han desarrollado diferentes tipos de bibliotecas. Un tipo es independiente de la columna vertebral, mientras que el otro es dependiente.

A continuación se define el concepto de rotámero y se describen los dos principales tipos de bibliotecas.

Un **rotámero** es una conformación de la cadena lateral de baja energía. Se sabe que los rotámeros no están uniformemente distribuídos, sino que se agrupan en regiones. Estos estudios dieron origen a las bibliotecas de rotámeros y el uso de las mismas para la predicción de estructuras.

La palabra rotámero (rotamer en inglés) viene de isómero rotacional (**rota**tional iso**mer** en inglés). Los rotámeros se obtienen usando proteínas de estructuras conocidas de alta calidad, obtenidas mediante el método experimental de cristalografía por rayos X.

Se encontró que hay una correlación entre las probabilidades de los ángulos de torsión de la cadena lateral y los valores ϕ y ψ de la columna vertebral (Dunbrack y Karplus, 1993). Esto da origen a dos tipos de bibliotecas de rotámeros, las independientes y las dependientes de la columna vertebral.

La **biblioteca de rotámeros independiente de la columna vertebral** contiene una colección de rotámeros por cada tipo de aminoácido con una probabilidad asociada a cada rotámero. Esta biblioteca es fácil de utilizar, pero al estudiar los nuevos métodos para el PSCPP (Krivov *et al.*, 2009; Wang *et al.*, 2008; Hsin *et al.*, 2007; Xu y Berger, 2006; Kingsford *et al.*, 2005; Peterson *et al.*, 2004; Chazelle *et al.*, 2004; Bower *et al.*, 1997) se nota una tendencia hacia el uso de bibliotecas dependientes de la columna vertebral.

En la Tabla VI se presenta la información básica de algunos registros de una biblio-

Tabla VI. Información básica de una biblioteca de rotámeros indepentiente de la columna vertebral.

Tipo de Aminoácido	χ_1	χ_2	χ_3	χ_4	$p(\chi_1, \chi_2, \chi_3, \chi_4)$
ARG	55.4	79.7	62.4	82.3	0.04
ARG	59.2	85.4	68.2	-166.2	0.07
...	...	...	...	...	...
VAL	65.5	n/a	n/a	n/a	7.40
VAL	175.9	n/a	n/a	n/a	73.90
VAL	-61.7	n/a	n/a	n/a	18.70

teca de rotámeros independiente de la columna vertebral. Se sabe que para aminoácidos muy grandes, como la arginina, se tienen hasta cinco ángulos de torsión para la cadena lateral. Para este caso, el ángulo χ_5 no se registra en las bibliotecas de rotámeros y al momento de utilizarse se establece como 180^o[1]. Para residuos pequeños como la valina. La tabla muestra también la probabilidad de ocurrencia para cada rotámero $p(\chi_1, \chi_2, \chi_3, \chi_4)$.

Cabe recordar que los aminoácidos glicina (GLY) y alanina (ALA) no tienen rotámeros, ya que no tienen ángulos de torsión en la cadena lateral (ver Figura 2, en ésta se muestra la estructura de cada aminoácido).

Las **bibliotecas de rotámeros dependientes de la columna vertebral** (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997; Dunbrack, 2002) contienen una colección de rotámeros por cada tipo de aminoácido para diferentes conformaciones de la columna vertebral; es un poco más compleja de utilizar, pero a pesar de ello, se obtienen mejores resultados con ellas (Bower *et al.*, 1997).

En la Tabla VII se muestra la información básica de algunos registros de una biblioteca de rotámeros dependiente de la columna vertebral. A diferencia de la biblioteca de rotámeros independiente, esta biblioteca contiene la información de la conformación de la columna vertebral, agrupando los rotámeros por el tipo de aminoácido y por los ángulos de torsión ϕ y ψ. Estos ángulos están en pasos de 10 grados, es decir, se tiene un conjunto de rotámeros para los residuos del aminoácido arginina cuyos ángulos de torsión están en los intervalos $-180 \leq \phi < -170$ y $-180 \leq \psi < -170$, otro conjunto para los residuos del aminoácido arginina cuyos ángulos de torsion están en

[1]Según lo indicado, en comunicación personal, por R. L. Dunbrack el 15 de febrero de 2010.

Tabla VII. Información básica de una biblioteca de rotámeros dependiente de la columna vertebral.

Tipo de Aminoácido	ϕ	ψ	χ_1	χ_2	χ_3	χ_4	p($\chi_1, \chi_2, \chi_3, \chi_4$)
ARG	-180	-180	55.4	79.7	62.4	82.3	0.002977
ARG	-180	-180	59.2	85.4	68.2	-166.5	0.006091
...	...	...	...	...	...	...	...
ARG	-180	-170	55.4	79.7	62.4	82.3	0.003496
ARG	-180	-170	59.2	85.4	68.2	-166.2	0.007153
...	...	...	...	...	...	...	...
VAL	180	180	65.6	n/a	n/a	n/a	0.333227
VAL	180	180	175.9	n/a	n/a	n/a	0.003134
VAL	180	180	-61.8	n/a	n/a	n/a	0.663638

$-180 \leq \phi < -170$ y $-170 \leq \psi < -160$, etc.

III.1.2 Función objetivo

Para definir un problema de optimización es necesario establecer la función que se requiere minimizar o maximizar, según sea el caso. En el PSCPP la función objetivo es un modelo aproximado de la energía libre de la proteína.

Para modelar fenómenos físicos es necesario identificar las variables que intervienen en el proceso. Los modelos que se generan pueden despreciar o incluir estas variables dependiendo del grado de complejidad y precisión que requiera la aplicación.

En el proceso de plegamiento se ven involucradas diversas interacciones fisico-químicas como: enlaces covalentes, puentes iónicos, puentes de hidrógeno, interacciones dipolo-dipolo, interacciones de Van der Waals. Para los enlaces covalentes hay otro tipo de potenciales que limitan los grados de libertad del enlace covalente. Estos potenciales están relacionados con: la longitud del enlace, el ángulo de los enlaces, los ángulos de torsión, entre otros (Steinbach, 2005a).

Los átomos que rodean a la proteína también intervienen en el proceso de plegado, como por ejemplo, las moléculas de agua crean puentes de hidrógeno con los residuos expuestos en la proteína.

A pesar de tener identificadas las variables que definen la energía de una proteína, estas variables no son fáciles de calcular debido al gran número de átomos que se ven

involucrados, así como a la dependencia entre algunas de estas variables.

Los cambios en temperatura, pH, o la presencia de otras moléculas y iones en el medio, pueden ocasionar cambios conformacionales importantes en la proteína.

En el PSCPP se desprecian muchas variables y se utilizan modelos simples que incluyen, únicamente, información de la proteína y no del medio. Esta aproximación es posible debido a que se requiere predecir la estructura de proteínas en medios controlados, es decir, con temperatura y pH preestablecidos, así como enlaces covalentes con longitudes y ángulos promedio.

En el modelado de otros fenómenos físicos, como el desplazamiento de un objeto, se definen las variables a considerar en el modelo y las que se pueden ignorar. Por ejemplo, un modelo sencillo puede incluir la posición, velocidad y acelaración en un instante dado, ignorando otras variables como la fricción, flujo del viento, etc. Si se incluye a la fricción, ésta se puede modelar como una constante o bien ser una función del tiempo, temperatura u otros factores. La inclusión o exclusión de variables en el modelo pueden afectar en la predicción. La aplicación es la que demanda el nivel de precisión del modelo.

En la predicción de estructuras se desea un nivel alto de precisión. Las limitantes que se tienen para medir algunas de las variables del modelo, como la entropía y las interacciones entre átomos del medio, son un obstáculo para utilizar un modelo más completo para esta aplicación. Para compensar el uso de modelos sencillos, se utiliza la información de estructuras conocidas para aumentar la calidad de la predicción de los modelos. Los modelos utilizados para el PSCPP se han estado mejorando paulatinamente, agregando un poco más de complejidad con el fin de aumentar la precisón de las predicciones.

También se ha intentado utilizar simulaciones basadas en dinámica molecular (Steinbach, 2005b) para predecir la estructura de proteínas. Estas simulaciones toman a la proteína en tiempo cero (t_0) y recalculan la posición, fuerza y velocidad de cada átomo para $t_i = t_{i-1} + \Delta t$, hasta que se llegue a un estado de equilibrio. Tienen la desventaja de ser muy lentas y costosas computacionalmente, y hasta el momento sólo se utilizan para moléculas pequeñas (Steinbach, 2005b).

Las funciones de energía más sencillas son aquellas en las que no se involucra el medio donde se encuentra la proteína, sino que sólo utilice la información de la misma para predecir su estructura. Estos modelos funcionan satisfactoriamente para predecir

la posición de los residuos que están en el interior de la molécula, pero las posiciones de los residuos expuestos al solvente son difíciles de predecir.

Los métodos desarrollados para el problema general de predicción de estructura se pueden clasificar en tres clases (Gu y Bourne, 2009): modelado homólogo (Sánchez y Sali, 1997), reconocimiento de plegado (Bryant y Altschul, 1995) y *ab initio* (Osguthorpe, 2000). En el modelado homólogo se utiliza la información de estructuras conocidas para la predicción de la estructura tridimensional de secuencias homólogas, es decir, aquellas que comparten 40% o más de identidad de secuencia. Para la predicción de estructuras de secuencias que no tienen homólogas con estructuras conocidas están los métodos de reconocimiento de plegado y *ab initio*. En los métodos *ab initio* se intentan utilizar funciones con términos que provengan de fenómenos físicos, como las fuerzas e interacciones antes mencionadas. En los otros tipos de métodos de modelado homólogo y reconocimiento de plegado se utilizan funciones que utilizan además de términos basados en propiedades físicas, términos basados en conocimiento, es decir, términos que se obtuvieron a través del análisis de estructuras conocidas, con parámetros optimizados.

Debido a que el PSCPP es parte del modelado homólogo, la función objetivo que se utiliza para este problema incluye términos basados en conocimiento previo. Los métodos actuales para el PSCPP utilizan funciones distintas, aún no hay un consenso de la función de energía que se debe utilizar para modelar este problema (Plaxco *et al.*, 2000; Gu y Bourne, 2009).

Uno de los términos más utilizados en las funciones de energía son aquellos que modelan las interacciones de Van der Waals, así como las probabilidades de los rotámeros seleccionados. Se ha demostrado experimentalmente que ambos términos tienen gran influencia en la predicción de la estructura tridimensional. Sin embargo, se ha incorporado más detalle a las funciones de energía para tratar de elevar la calidad de las predicciones como los relacionados con: interacciones electrostáticas, solvatación, puentes de hidrógeno, choques interatómicos, orientación de las cadenas laterales, etc.

En la Ecuación 5 se presenta una función de energía sencilla que aproxima el potencial generado por las interacciones de Van der Waals. Esta función se utiliza en varios métodos (Yanover *et al.*, 2008; Lu *et al.*, 2008a,b; Zhang *et al.*, 2008; Krivov *et al.*, 2009; Liang y Grishin, 2002; Canutescu *et al.*, 2003; Jain *et al.*, 2006; Wang *et al.*, 2005; Peterson *et al.*, 2004; Kingsford *et al.*, 2005; Hsin *et al.*, 2007).

$$E = \sum_i E_{lib}(i) + \sum_{a,b} E(a,b) \tag{5}$$

$$E_{lib}(i) = -K \log \frac{p(r_i|R,\phi,\psi)}{p(r_i = 1|R,\phi,\psi)} \tag{6}$$

$$E(a,b) = \begin{cases} 0 & d(a,b) \geq r(a) + r(b) \\ 10 & d(a,b) \leq 0.8254(r(a) + r(b)) \\ 57.273(1 - \frac{d(a,b)}{r(a)+r(b)}) & otro \end{cases} \tag{7}$$

donde E es la energía libre de una conformación, i es el i-ésimo residuo, a y b representan a átomos de radio $r(a)$ y $r(b)$ con centros separados a una distancia $d(a,b)$.

La distancia $d(a,b)$ se obtiene fácilmente, ya que se conocen las coordenadas de los átomos a y b. Los radios $r(a)$ y $r(b)$ son los radios de Van der Waals, para los cuales se utilizan los siguientes valores: carbono, 1.6Å; oxígeno, 1.3Å; nitrógeno, 1.3Å; y azúfre, 1.7Å. Estos valores son aproximados, y se han hecho estudios (Peterson *et al.*, 2004; Liang y Grishin, 2002) donde se utilizan diferentes valores para los radios de los átomos, variando la calidad de las soluciones dependiendo del conjunto que se utilice.

La constante $K(= 3)$ se optimizó con un conjunto de prueba, $p(r_i|R,\phi,\psi)$ es la probabilidad del rotámero seleccionado para el residuo i que se obtiene de la biblioteca de rotámeros dependiente de la columna vertebral (R). La probabilidad $p(r_i = 1|R,\phi,\psi)$ es la suma de las probabilidades de todos los rotámeros dados el tipo de aminoácido, y los ángulos ϕ y ψ, es decir, se normaliza el valor $p(r_i|R,\phi,\psi)$.

Dada una estructura tridimensional, evaluar la función de energía aquí mostrada tiene una complejidad de $O(n^2)$, donde n es la cantidad de residuos de la proteína. En el PSCPP se requiere realizar este cálculo en múltiples ocasiones, (el problema pertenece a la clase *NP-difícil* (Akutsu, 1997)).

Recapitulando, para el PSCPP se tiene un modelo aproximado de la energía libre, la cual se quiere minimizar. Esto debido a las leyes de la termodinámica, que establecen que un sistema toma el estado de mínima energía libre. En mecánica estadística, la probabilidad de un estado i está dada por la Ecuación 8, conocida como la distribución de Boltzman.

$$p_i = \frac{e^{-\frac{E_i}{k_B T}}}{\sum_{j=1}^{l} e^{-\frac{E_j}{k_B T}}} \tag{8}$$

donde p_i es la probabilidad del estado i, E_i es la energía del estado i, k_B es la constante de Boltzman, T es la temperatura en grados Kelvin y l es la cantidad de niveles de energía en el sistema.

Se puede observar entonces, que para determinar entre dos estados i y j aquel que tiene mayor probabilidad, se puede descartar el denominador y comparar únicamente los numeradores de la ecuación de la distribución de Boltzman, que con manipulación algebraica queda, $p_i > p_j$ si y sólo si $E_i < E_j$. Esto da la base para aproximar la energía, y encontrar el estado que la minimice, o que es lo mismo, que aumente su probabilidad.

En resumen, los métodos que resuelven el PSCPP, utilizan modelos aproximados de la energía libre, con términos basados en propiedades físicas y otros obtenidos estadísticamente. La salida al problema es la conformación de las cadenas laterales que minimice la energía libre.

En (Lu *et al.*, 2008b) se presenta un análisis sobre la influencia de los términos de la función de energía utilizada para la predicción. Este método utiliza la técnica de recocido simulado (Kirkpatrick *et al.*, 1983), la cual tiene la ventaja de modificar su función objetivo fácilmente sin hacer grandes cambios en el algoritmo. La función de energía utilizada está dada por la siguiente expresión.

$$E_{total} = E_{rot} + w_{vdw}E_{vdw} + w_{orient}E_{orient} + w_{solvation}E_{solvation} \tag{9}$$

donde E_{rot} es el término que involucra las probabilidades de los rotámeros, E_{vdw} representa el potencial 6-12 de Lennard-Jones (Gray *et al.*, 2003), E_{orient} es el potencial de empaquetamiento OPUS-PSP (Lu *et al.*, 2008a) y $E_{solvation}$ es un término de solvatación. Los pesos $w_{vdw} = 1.0$, $w_{orient} = 0.15$ y $w_{solvation} = 0.1$ fueron optimizados utilizando un conjunto pequeño de proteínas de alta resolución.

Lu *et al.* (2008b) modificaron la función de energía original (Ecuación 9), eliminando un término a la vez y comparando la calidad del método con las nuevas funciones. En la Tabla VIII se presentan los resultados de dicha comparación incluyendo todos los residuos e incluyendo únicamente los resiudos internos[2].

En base a esta comparación se puede observar que el término que más influye en la predicción es E_{vdw}, ya que al eliminarlo de la función de energía, disminuye la precisión en mayor medida que si se eliminan los demás términos. Sin embargo, todos los términos

[2]Lu *et al.* (2008b) definen residuos internos como aquellos cuyo porcentaje de superficie expuesta al solvente es menor al 17%.

Tabla VIII. Resultados presentados en (Lu *et al.*, 2008b) para notar la influencia de cada término en la predicción global.

	Todos los residuos		Residuos internos	
	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$
OPUS-Rota	**89.0**	**79.1**	**94.5**	**88.7**
sin E_{vdw}	81.1	68.5	85.0	75.3
sin E_{rot}	83.8	69.0	92.6	83.7
sin E_{orient}	88.3	77.2	93.9	87.2
sin $E_{solvation}$	88.6	78.6	94.0	88.5

aportan positivamente a la calidad de las predicciones, esto se analiza más adelante cuando se comparen métodos con funciones de energía diferentes.

El PSCPP pertenece a la clase NP-difícil (Akutsu, 1997), pero se han desarrollado diferentes métodos que hacen el problema manejable, algunos basados en teoría de grafos, en técnicas de ramificación y acotamiento ($branch - and - bound$), y algunas heurísticas como recocido simulado. A continuación se hará mención de algunas propuestas para atacar el PSCPP.

III.1.3 Definición matemática del PSCPP

Primero se definen las variables que se van a utilizar en las secciones posteriores.

Aminoácidos

Sea L el conjunto que representa a los 20 aminoácidos estándar. Entonces

$$\begin{aligned}\mathbf{L} &= \{\text{ALA, CYS, ASP, GLU, PHE, GLY, HIS, ILE, LYS, LEU, MET, ASN,} \\ &\quad \text{PRO, GLN, ARG, SER, THR, VAL, TRP, TYR}\}\end{aligned}$$

Biblioteca de rotámeros

Sea **RL** el conjunto que representa a la biblioteca de rotámeros. Entonces

$$\begin{aligned}\mathbf{RL} &= \{R_{ALA}, R_{CYS}, R_{ASP}, R_{GLU}, R_{PHE}, R_{GLY}, R_{HIS}, R_{ILE}, R_{LYS}, R_{LEU}, R_{MET} \\ &\quad R_{ASN}, R_{PRO}, R_{GLN}, R_{ARG}, R_{SER}, R_{THR}, R_{VAL}, R_{TRP}, R_{TYR}\} \\ R_x &= (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_{|R_x|})\end{aligned}$$

donde R_x, para un $x \in \mathbf{L}$, es la biblioteca de rotámeros para el aminoácido tipo x y $\mathbf{r}_i$ que está en R_x es el i-ésimo rotámero para el aminoácido tipo x.

Secuencia de aminoácidos

Sea A la variable de entrada que representa la secuencia de aminoácidos, ℓ la longitud de esta secuencia, entonces

$$A = (a_1, a_2, a_3, \ldots, a_\ell)$$

donde $a_i \in \mathbf{L}$ es el tipo de aminoácido del i-ésimo residuo de la cadena.

Coordenadas de la columna vertebral

Sea N, CA, C y O las variables de entrada que representan la secuencia de coordenadas de los átomos (nitrógeno, carbono-α, carbono y oxígeno, respectivamente) de la columna vertebral de la proteína, entonces

$$\begin{aligned} N &= (n_1, n_2, n_3, \ldots, n_\ell) \\ CA &= (ca_1, ca_2, ca_3, \ldots, ca_\ell) \\ C &= (c_1, c_2, c_3, \ldots, c_\ell) \\ O &= (o_1, o_2, o_3, \ldots, o_\ell) \\ n_i, ca_i, c_i, o_i &\in \Re^3 \; \forall \; 1 \leq i \leq \ell \end{aligned}$$

Solución del PSCPP

A continuación se describe la forma en la que se representa una solución en el algoritmo. Para este problema, una solución S es una secuencia de enteros que indican los índices de los rotámeros que se seleccionaron, uno por cada residuo. Es decir,

$$S = (s_1, s_2, s_3, \ldots, s_\ell)$$

donde $1 \leq s_i \leq |R_{a_i}|$.

El problema de empaquetamiento de la cadena lateral en proteínas (PSCPP)

El PSCPP tiene como entrada A, que es la secuencia de aminoácidos; N, C, CA y O, que son las coordenadas de la columna vertebral; y **RL** que es la biblioteca de rotámeros. El PSCPP consiste en encontrar S^*, que se define a continuación.

$$S^* = \arg\min E(S) \tag{10}$$

donde $E(\cdot)$ es la función objetivo propuesta para modelar la energía de la conformación predicha.

III.2 Trabajo previo

El PSCPP ha sido tema de investigación por muchos años. Se han desarrollado diferentes métodos que lo resuelven, tanto exactos como heurísticas. Los cambios que han sufrido estos métodos son en tres aspectos principalmente: biblioteca de rotámeros, función de energía y método de búsqueda.

Se han desarrollado varias bibliotecas de rotámeros para mejorar la calidad de los métodos que las utilizan (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997; Dunbrack, 2002; Krivov *et al.*, 2009). Las bibliotecas se obtienen estadísticamente, utilizando proteínas con estructuras conocidas de alta resolución; sin embargo, las primeras bibliotecas se desarrollaron utilizando conjuntos muy pequeños, ya que la cantidad de proteínas con estructuras conocidas era muy reducida.

Con el paso de los años y los múltiples esfuerzos de obtener las estructuras de más y más proteínas, se tienen conjuntos más grandes y diversos con los que se han podido generar mejores bibliotecas de rotámeros. Los nuevos métodos para el PSCPP (Krivov *et al.*, 2009; Wang *et al.*, 2008; Hsin *et al.*, 2007; Xu y Berger, 2006; Kingsford *et al.*, 2005; Peterson *et al.*, 2004; Chazelle *et al.*, 2004; Bower *et al.*, 1997) utilizan las nuevas bibliotecas y han generado mejores resultados en sus predicciones. Pero hay que considerar que a la par se han modificado también las funciones de energía (Krivov *et al.*, 2009; Yanover *et al.*, 2008; Lu *et al.*, 2008a; Grigoryan *et al.*, 2007; Peterson *et al.*, 2004; Liang y Grishin, 2002), por lo que ambos aspectos han influido positivamente en mejorar las predicciones.

El desarrollo de nuevos métodos de búsqueda han permitido que el tiempo de eje-

cución de los mismos disminuya y se puedan utilizar como parte del modelado homólogo para la predicción de estructuras. A continuación se presentan algunos métodos que se han propuesto para resolver el PSCPP, de los cuales, el "Dead-End Elimination" (DEE) es uno de los pioneros y se ha incorporado como parte de otros nuevos métodos.

III.2.1 SCWRL: Side-Chain placement With Rotamer Library

El método SCWRL tiene una larga historia. Desde su primera versión (Bower *et al.*, 1997), el método se ha ido mejorando, cambiando la biblioteca de rotámeros, la función de energía y el algoritmo de búsqueda.

La última versión es el SCWRL4 (Krivov *et al.*, 2009), que hasta el momento tiene 3919[3] licencias concedidas. Su versión previa, SCWRL3, también se utilizó ampliamente (Krivov *et al.*, 2009) para resolver el PSCPP como parte de otros métodos de predicción de estructura.

Las características que hacen que estos métodos sean tan utilizados son: exactitud, velocidad y usabilidad[4].

SCWRL3

El SCWRL3 (Canutescu *et al.*, 2003) es un método exacto, ya que asegura obtener la solución óptima de la función de energía propuesta. Se basa en el método DEE (*dead-end elimination*) que establece una forma de eliminar rotámeros de alta energía que no son parte de la solución, para reducir el espacio de búsqueda. Además utiliza técnicas de teoría de grafos y de ramificación y acotamiento para encontrar el óptimo a partir de los rotámeros restantes. La utilización de estas técnicas permitieron que el SCWRL3 sea un método exacto, ya que descompone el problema en subproblemas cuyas soluciones son independientes, disminuyendo la complejidad computacional del algoritmo.

Este método demostró superar a los previos en cuanto rapidez y exactitud, además de tener una interfaz fácil de adaptar como parte de otros métodos, teniendo una buena usabilidad.

[3]http://dunbrack.fccc.edu/scwrl4/index.php - 16 de noviembre de 2010

[4]Usabilidad: conjunto de atributos relacionados con el esfuerzo necesario para su uso, y con la evaluación individual de tal uso, por una conjunto de usuarios explícito o implícito (Bevan, 2001).

La función de energía usada por este método incorpora términos en donde se involucra la probabilidad de los rotámeros seleccionados y una penalización por choques interatómicos utilizada desde la primer versión del SCWRL.

SCWRL4

El SCWRL4 (Krivov *et al.*, 2009) mejoró (usando las medidas de calidad mencionadas en la Sección III.3) al SCWRL3. Los cambios del SCWRL4 fueron significativos, entre los que destacan los siguientes.

1. Se generó una biblioteca de rotámeros nueva, usando nuevas técnicas que proponen mejorar las bibliotecas anteriores.

2. Incorporan la idea de subrotámeros, para permitir cierta flexibilidad en la ubicación de las cadenas laterales.

3. Modificaron la función de energía. El término que representaba sólo la fuerza repulsiva, se transformó para aumentar su semejanza con el potencial de van der Waals, el cual incluye una parte de atracción. Este término se obtiene de las distancias entre pares de átomos. Se agregó en la función de energía un término para indicar el potencial de los puentes de hidrógeno, ya que se sabe que los puentes de hidrógeno son parte importante de la estabilidad de la proteína y de su plegado.

4. Como parte del algoritmo, este método utiliza la técnica de "Tree-decomposition of a graph" para dividir el problema en subproblemas independientes.

5. A pesar de usar el DEE y las técnicas para dividir el problema en subproblemas más pequeños, el SCWRL3 podría no converger para ciertos casos. Por lo tanto, el SCWRL4 tomó en consideración esto y establece un tiempo máximo de búsqueda, de tal manera que si no se encuentra el óptimo para cuando transcurre el tiempo máximo, se devuelve una aproximación a éste.

Es posible darse cuenta que el SCWRL4 intenta mejorar la calidad de sus soluciones con los primeros tres cambios, y los últimos dos son para no afectar, o bien disminuir, el tiempo de búsqueda.

SCWRL4 demostró experimentalmente mejorar la calidad que el SCWRL3 y reemplazar este último por el SCWRL4 es fácil, ya que tienen la misma interfaz.

La usabilidad de estos métodos es muy buena, ya que mantiene la numeración original de los aťomos y residuos, comparada a otros métodos, como el OPUS-Rota (Lu *et al.*, 2008b), que cambian la numeración de los átomos y residuos, haciendo más difícil la comparación de la predicción contra el modelo original.

El SCWRL4 toma como entrada un archivo en formato PDB dónde se encuentran las coordenadas de los átomos pesados de la columna vertebral, y genera como salida otro archivo en formato PDB con las coordenadas de los átomos de las cadenas laterales, además de las que se dieron como entrada. Cabe mencionar, que por facilidad se le puede dar como entrada el archivo original de una estructura que se desea predecir, ya que recalcula todas las coordenadas, considerando sólo las coordenadas de los átomos de la columna vertebral. Este programa está disponible para fines académicos[5].

III.2.2 OPUS-Rota

El método OPUS-Rota es un método basado en la técnica de recocido simulado que aproxima el PSCPP. Este método utiliza una función de energía que modela las interacciones de Van der Waals, el término de las probabilidades de los rotámeros utilizados, un término de solvatación y un término de orientación (ver Ecuación 9).

En este método, lo novedoso es la función de energía. El método de búsqueda ya se había utilizado con anterioridad (Peterson *et al.*, 2004; Liang y Grishin, 2002; Jain *et al.*, 2006; Wang *et al.*, 2005; Xiang y Honig, 2001); sin embargo, los términos de solvatación y orientación son nuevos.

El tiempo de ejecución reportado es comparable al del SCWRL 3 (Canutescu *et al.*, 2003), que hasta el momento era el más rápido. La biblioteca de rotámeros utilizada fue la biblioteca de rotámeros dependiente de la columna vertebral, desarrollada por Dunbrack y Cohen (1997).

Las ventajas que ofrece este método es la flexibilidad en la función de energía utilizada. Como se mencionó anteriormente, se hizo un análisis de la función de energía utilizada por este método y el aporte de cada término para la calidad de las soluciones (ver Tabla VIII). Gracias a que el método está basado en recocido simulado, se pudo

[5]http://dunbrack.fccc.edu/scwrl4/index.php

modificar la función de energía fácilmente sin afectar el algoritmo.

La misma idea se puede utilizar para agregar términos a la función de energía y evaluar su desempeño. Esto puede ayudar para refinar la función que se utiliza para modelar el plegado de las proteínas.

Algunos aspectos a modificar en este método que pudieran resultar en mejoras son: función de vecindario, función de energía, parámetros del recocido simulado, biblioteca de rotámeros. La modificacion más sencilla sería ajustar los parámetros del recocido simulado, aunque este cambio no asegura una mejora. Sin embargo, actualizar la biblioteca de rotámeros representaría una gran ventaja, ya que ha surgido una nueva biblioteca a partir de la fecha de publicación del método.

Al igual que el SCWRL4, el OPUS-Rota toma como entrada un archivo en formato PDB y genera como salida otro archivo en formato PDB con las coordenadas faltantes. Una desventaja del OPUS-Rota es que no calcula las coordenadas del C-β, por lo que se tienen que calcular previo al uso del OPUS-Rota. Como se verá en la sección A.1.6, el C-β es un átomo redundante de la cadena lateral, por lo que determinar las coordenadas de este átomo no requiere optimización. Se debe tener cuidado al comparar el modelo resultante con el original, ya que las numeraciones de los átomos y residuos pueden cambiar.

En resumen, para usar el OPUS-Rota es recomendable procesar previamente los archivos originales de las estructuras que se desean probar, borrar todos los átomos de las cadenas laterales, y posteriormente calcular las coordenadas de los C-β.

Este programa está disponible para fines académicos[6], y sólo está disponible para ejecutarse en sistemas operativos Linux con arquitectura x86.

En conclusión, los métodos actuales están limitados por dos aspectos. Primero, las funciones de energía que se han desarrollado son aproximaciones, y el mínimo global de estas funciones no siempre representa la conformación que se está buscando. Segundo, las bibliotecas de rotámeros se obtienen estadísticamente basándose en estructuras conocidas, por lo tanto, carece de otras conformaciones que si bien pueden ser de alta energía la mayoría de las veces, se pueden presentar en la naturaleza bajo ciertas condiciones.

Por lo tanto, los métodos para el PSCPP deben tener la flexibilidad suficiente para modificar estos dos aspectos fácilmente y así evolucionar junto con el desarrollo de nuevas bibliotecas de rotámeros y funciones de energía.

[6]http://sigler.bioch.bcm.tmc.edu/MaLab/soft/opus_rota.html

III.3 Medidas de calidad

Existen métodos exactos y heurísticas que resuelven el PSCPP Se llaman métodos exactos a aquellos que encuentran el mínimo global de la función objetivo que se está utilizando. Sin embargo, la función objetivo es una aproximación de la energía libre, por lo tanto, el óptimo global puede o no corresponder al estado real de mínima energía libre de la proteína.

Para ambos tipos de métodos, exactos y heurísticas, se reporta la calidad de las soluciones. Para realizar experimentos y saber qué tan bien funciona un método se buscan conjuntos de prueba de proteínas con estructuras conocidas. Se eliminan los átomos de las cadenas laterales de los archivos originales, y se utilizan los métodos para la predicción de las cadenas laterales. Una vez terminada la predicción se utilizan diferentes tipos de medidas para comparar las estructuras.

Hay medidas diseñadas para comparar estructuras, como RMSD (Gu y Bourne, 2009), TM-score (Zhang y Skolnick, 2004), entre otras, las cuales consideran a todos los átomos. En el PSCPP se utilizan otros tipos de medidas como exactitud absoluta y condicional (Krivov *et al.*, 2009). Éstas se enfocan únicamente en las cadenas laterales. El RMSD también lo utilizan algunos autores para presentar sus resultados.

A continuación se describen las medidas RMSD, exactitud absoluta y exactitud condicional.

III.3.1 Desviación raíz media cuadrática (RMSD)

El RMSD se utiliza como una medida de diferencia entre estructuras. En la Ecuación 11 se muestra la fórmula para obtener el RMSD entre un par de estructuras.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{11}$$

donde δ_i es la distancia entre el par i de los N pares de átomos de las estructuras.

Esta medida se puede utilizar para comparar diferentes proteínas. Para hacerlo, se tiene que pasar por un proceso de alineamiento de estructuras para que se puedan encontrar los pares de átomos equivalentes en las estructuras.

Para esta aplicación, se conoce de antemano los pares equivalentes, por lo que no es necesaria la etapa de alineamiento de estructuras. El cálculo de esta medida para un

par de estructuras superimpuestas de n residuos es de $O(n)$.

Para el PSCPP se desarrollaron otras medidas que usan únicamente la información de las cadenas laterales para su evaluación. A continuación se desribe la medida exactitud absoluta, la cual es la más utilizada para reportar la calidad de los métodos para el PSCPP.

III.3.2 Exactitud absoluta

La medida exactitud absoluta (AA[7]) se utiliza para evaluar la calidad de las soluciones de los métodos que resuelven el PSCPP. Se comparan los modelos orginales contra las predicciones generadas por estos métodos. La única diferencia entre el modelo original y la predicción es en las cadenas laterales. Los modelos originales y las predicciones de los métodos que resuelven el PSCPP están dados en archivos de texto en formato PDB que contiene las coordenadas 3D de todos los átomos de la proteína, tanto de la columna vertebral como de las cadenas laterales.

Para poder hacer la comparación es necesario que el modelo original contenga la información de las cadenas laterales. Esto se debe prever al momento de seleccionar el conjunto de prueba para realizar el análisis de los métodos.

Con las coordenadas de los átomos de las cadenas laterales se pueden obtener los ángulos de torsión de cada residuo, tanto del modelo original como de la predicción. Para obtener la medida AA se utilizan únicamente los valores de los ángulos, y no las coordenadas como es el caso del RMSD.

Los residuos tienen diferentes tamaños, y por ende, diferente número de ángulos de torsión en la cadena lateral. En la Tabla IX se muestra la cantidad de ángulos de torsión de la cadena lateral por cada residuo.

De manera general, la medida AA es el porcentaje de ángulos correctos respecto al total de residuos, es decir, se calcula el total de residuos cuyos ángulos de torsión predichos sean correctos, y se divide entre el total de residuos en la proteína. Sin embargo, como ya se dijo anteriormente, no todos los residuos tienen ángulos de torsión de la cadena lateral, y además varía la cantidad de ángulos según el residuo, por lo que se tienen que considerar esto al momento de evaluar las medidas.

Otro problema es que se van a comparar ángulos de torsión, los cuales están en

[7]AA: Absolute Accuracy

Tabla IX. Cantidad de ángulos de torsión de la cadena lateral por aminoácido.

Aminoácido (código de 1 letra)	Ángulos de torsión de la cadena lateral	Aminoácido (código de 1 letra)	Ángulos de torsión de la cadena lateral
A	0	M	3
C	1	N	2
D	2	P	0
E	3	Q	3
F	2	R	5
G	0	S	1
H	2	T	1
I	2	V	1
K	4	W	2
L	2	Y	2

un espacio continuo, y errores de cálculo numérico pueden afectar a las medidas. Este problema se resuelve tomando un margen de error para el cual aún se puede considerar un ángulo como correcto. Este valor varía según el autor del análisis; en especial, la mayoría de los trabajos reportan esta medida con un margen de 40^o (Lu *et al.*, 2008b; Canutescu *et al.*, 2003; Peterson *et al.*, 2004; Liang y Grishin, 2002; Jain *et al.*, 2006; Wang *et al.*, 2005; Xiang y Honig, 2001; Krivov *et al.*, 2009). Es decir, se considera como ángulo correcto aquel que esté alejado 40^o o menos del ángulo original.

De manera particular, las medidas reportadas son $\chi_1(\%)$, $\chi_{1+2}(\%)$, $\chi_{1+2+3}(\%)$, $\chi_{1+2+3+4}(\%)$ y $\chi_{1+2+3+4+5}(\%)$. En las Ecuaciones 12 - 16 se presentan las medidas antes mencionadas.

$$\chi_1(\%) = \frac{c_1}{n - n_A - n_G} \tag{12}$$

$$\chi_{1+2}(\%) = \frac{c_{1+2}}{n - n_A - n_G - n_C - n_S - n_T - n_V} \tag{13}$$

$$\chi_{1+2+3}(\%) = \frac{c_{1+2+3}}{n_E + n_K + n_M + n_Q + n_R} \tag{14}$$

$$\chi_{1+2+3+4}(\%) = \frac{c_{1+2+3+4}}{n_K + n_R} \tag{15}$$

$$\chi_{1+2+3+4+5}(\%) = \frac{c_{1+2+3+4+5}}{n_R} \tag{16}$$

donde n es la longitud de la cadena y n_X es la cantidad de residuos cuyo código de una letra es X. El parámetro c_1 es la cantidad de residuos predichos que tienen el ángulo de torisión χ_1 a 40^o o menos del ángulo original. El parámetro c_{1+2} es la cantidad de residuos predichos que tienen los ángulos de torsión χ_1 y χ_2 alejados a lo más 40^o del original. De la misma manera, los parámetros c_{1+2+3}, $c_{1+2+3+4}$ y $c_{1+2+3+4+5}$ representan la cantidad de residuos predichos que tienen correctos los ángulos desde χ_1 hasta χ_3, χ_4 y χ_5, respectivamente.

En esta medida se va acumulando el error, por ejemplo, si se tiene un residuo del tipo arginina, cuyo ángulo χ_1 es incorrecto (está alejado más de 40^o del ángulo original) este residuo se omite en c_1, c_{1+2}, c_{1+2+3}, $c_{1+2+3+4}$ y $c_{1+2+3+4+5}$ a pesar de que alguno de los otros ángulos χ_2, χ_3, χ_4 o χ_5 sean correctos.

Es importante mencionar que los ángulos de torsión están en el intervalo (-180,180]. Para calcular la distancia del ángulo original (χ_i^{ori}) respecto al predicho (χ_i^{pre}) es necesario hacer algunos ajustes debido al intervalo en el que se encuentran los ángulos. Una simple resta (Ecuación 17) no es suficiente, ya que se tiene que considerar el cuadrante en el que se encuentran los ángulos. En la Ecuación 18 se muestra una manera sencilla de calcular la diferencia real entre los ángulos χ_i^{ori} y χ_i^{pre}.

$$\xi_i = \max(\chi_i^{ori}, \chi_i^{pre}) - \min(\chi_i^{ori}, \chi_i^{pre}) \quad (17)$$

$$\delta_i = \min(\xi_i, 360 - \xi_i) \quad (18)$$

Ahora bien, si $\delta_i \leq 40^o$ se considera al ángulo χ_i^{pre} como correcto.

El único aminoácido que tiene χ_5 es la arginina. La cadena lateral de la arginina es muy larga, y además es polar, por lo que es propenso a encontrarse en la superficie de las proteínas, ocasionando que la cadena lateral esté en constante movimiento y sea difícil obtener las coordenadas de los átomos que están más alejados de la columna vertebral, inclusive con los métodos experimentales de cristalografía por rayos X y resonancia magnética nuclear. Los métodos que resuelven el PSCPP, entonces, no predicen el ángulo χ_5 para los residuos tipo arginina, en su lugar, le asignan un valor fijo ($\chi_5 = 180^o$)[8]. Con esto se elimina la necesidad de reportar medidas para este ángulo ($\chi_{1+2+3+4+5}(\%)$).

[8]Según lo indicado, en comunicación personal, por R. L. Dunbrack el 15 de febrero de 2010.

Si la diferencia entre un par de ángulos correspondientes está a $40^o + \epsilon$ ($\epsilon > 0$), éste se considera como incorrecto. Ésta es una desventaja de utilizar esta medida; pero aún así se ha utilizado para poder comparar los nuevos métodos con los anteriores, ya que es la manera en la que se ha reportado la calidad de los previos.

Ahora bien, se puede tener otro caso en el que dos soluciones (B y B') tengan las mismas medidas de calidad absoluta. En este caso, los ángulos de la primera coinciden casi perfectamente con la original, mientras que los ángulos de la segunda difieren por un margen mucho mayor, pero sin llegar a 40^o. Bajo las medidas de exactitud absoluta, ambas soluciones son igual de buenas, pero vemos entonces, que hay una necesidad de utilizar otra medida donde la primera resulte obviamente mejor que la segunda. Esto se discute en el Capítulo V.

Otra medida muy similar es la exactitud condicional, que utiliza también la información de los ángulos para evaluar las predicciones. A continuación se describe la medida exactitud condicional.

III.3.3 Exactitud condicional

La exactitud condicional (CA[9]) es muy similar a la medida AA. Ambas se utilizan para medir la calidad de las predicciones generadas por métodos que resuelven el PSCPP. Utilizan únicamente la información de los ángulos de torsión de la cadena lateral para su evaluación.

Las medidas CA son: $\chi_1(\%)$, $\chi_2(\%)$, $\chi_3(\%)$, $\chi_4(\%)$ y $\chi_5(\%)$; sin embargo, $\chi_5(\%)$ no es reportada debido a que el ángulo de torsión χ_5 para los residuos tipo arginina es fijo, y no se predice. En las Ecuaciones 19 - 23 se presentan las medidas aquí mencionadas.

$$\chi_1(\%) = \frac{c_1}{n - n_A - n_G} \tag{19}$$

$$\chi_2(\%) = \frac{c_{1+2}}{c_1 - c_{(1)C} - c_{(1)S} - c_{(1)T} - c_{(1)V}} \tag{20}$$

$$\chi_3(\%) = \frac{c_{1+2+3}}{c_{(1+2)E} + c_{(1+2)K} + c_{(1+2)M} + c_{(1+2)Q} + c_{(1+2)R}} \tag{21}$$

$$\chi_4(\%) = \frac{c_{1+2+3+4}}{c_{(1+2+3)K} + c_{(1+2+3)R}} \tag{22}$$

$$\chi_5(\%) = \frac{c_{1+2+3+4+5}}{c_{(1+2+3+4)R}} \tag{23}$$

[9]CA: Conditional Accuracy.

donde $c_{(1)X}$ representa la cantidad de residuos cuyo código de una letra es X y ángulo de torsión χ_1 se predijo correctamente, es decir, que está alejado 40^o o menos del ángulo original. El parámetro $c_{(1+2)X}$ representa la cantidad de residuos cuyo código de una letra es X, y los ángulos de torsión χ_1 y χ_2 se predijeron correctamente. De la misma manera $c_{(1+2+3)X}$ y $c_{(1+2+3+4)X}$ son los residuos tipo X, y que sus ángulos desde χ_1 hasta χ_3 y χ_4, respectivamente, son correctos. Un aspecto interesante es que la medida $\chi_1(\%)$ se utiliza tanto para AA como para CA.

La medida más utilizada es el AA, aunque es interesante hacer el análisis con las demás medidas. Los resultados obtenidos con estas medidas han mostrado mejoría con el paso de los años, a pesar de que los ángulos χ_3 y χ_4 son difíciles de predecir. Esto se debe a que la cantidad de residuos que presentan estos ángulos es reducida comparada al total de residuos, y no se puede hacer un análisis con poca información. Además, los residuos que presentan estos ángulos son arginina, metionina, lisina, ácido glutámico y glutamina, los cuáles son polares y tienen preferencia por ubicarse en la superficie de las estructuras, por lo que tienen mucha flexibilidad en esas regiones y las restricciones son pocas para limitar la predicción.

En general, se han utilizado, principalmente, a las medidas $\chi_1(\%)$ y $\chi_{1+2}(\%)$ para reportar la calidad de las predicciones de los métodos que resuelven el PSCPP.

Con lo anterior se ha definido el PSCPP, y a continuación se define el problema que se resuelve en el presente trabajo.

III.4 Nuestro problema

Nuestro problema consiste en comparar los métodos SCWRL4 y OPUS-Rota, que según la literatura, son los mejores que aproximan el PSCPP. Para esto es necesario seleccionar un conjunto de prueba que sea representativo de las diferentes clases del SCOP y EC, ya que se intenta buscar una relación entre estas clases y la calidad de los métodos.

En este problema se tiene como entrada dos algoritmos (SCWRL4 y OPUS-Rota) y los casos de prueba; y como salida, el método con mejor exactitud.

Anteriormente se han desarrollado diferentes métodos, que se han comparado entre ellos utilizando los mismos conjuntos de pruebas reportados desde hace años, los cuales son muy pequeños y varias estructuras de estos son ahora obsoletas.

Los métodos SCWRL4 y OPUS-Rota son métodos con características diferentes.

Ambos métodos se han comparado anteriormente con el método SCWRL3 (Lu *et al.*, 2008b; Krivov *et al.*, 2009), pero no entre ellos.

Nuestro trabajo consiste, en encontrar un conjunto de pruebas representativo y posteriormente comparar estadísticamente estos métodos utilizando las medidas de calidad ya existentes.

El SCWRL4 utilizó 379 proteínas en su conjunto de pruebas, obteniendo 89.3% en la medida $\chi_1(\%)$. Es conveniente notar que sólo se están considerando los residuos cuya densidad de electrones esté en los percentiles del 25 al 100.

El OPUS-Rota utilizó un conjunto de pruebas de 65 proteínas, obteniendo 89% en la medida $\chi_1(\%)$.

Estas medidas están cercanas, pero aún así no se pueden comparar directamente ya que los conjuntos son diferentes y no se utilizaron los mismos criterios para decidir qué residuos utilizar para el análisis y cuáles no.

En el siguiente capítulo se describe cómo se obtiene el conjunto de pruebas que se propone, así como los resultados de la comparación de los métodos utilizando dicho conjunto.

Capítulo IV

COMPARACIÓN DE MÉTODOS

IV.1 Introducción

En este capítulo se presentan los experimentos, resultados y análisis realizados con el objetivo de comparar los mejores métodos desarrollados hasta el momento que aproximan el problema conocido como empaquetamiento de la cadena lateral en proteínas (PSCPP).

Los mejores métodos se seleccionaron basándose en las medidas de calidad mencionadas en la sección III.3.2. Las medidas de exactitud absoluta son las más usadas entre todos los métodos que aproximan el PSCPP (Lu *et al.*, 2008b; Canutescu *et al.*, 2003; Peterson *et al.*, 2004; Liang y Grishin, 2002; Jain *et al.*, 2006; Wang *et al.*, 2005; Xiang y Honig, 2001; Krivov *et al.*, 2009), por tal motivo se toman como punto de referencia para la selección de los métodos.

IV.2 Materiales y métodos

Los métodos que aquí se comparan son SCWRL4 (Krivov *et al.*, 2009) y OPUS-Rota (Lu *et al.*, 2008b), ya que son los que reportan las medidas de exactitud absoluta más altas en la actualidad. Ambos métodos difieren en los tres aspectos principales: función objetivo, biblioteca de rotámeros y método de búsqueda.

El método SCWRL4 es una heurística determinística. Es la cuarta versión del método SCWRL que era, en su tercera versión, SCWRL3 (Canutescu *et al.*, 2003), un método exacto, ya que asegura obtener el óptimo global de la función de energía ahí propuesta.

El SCWRL4 incorpora ideas de otros métodos (Xu, 2005), y presenta una mejora en calidad contra el método SCWRL3. El algoritmo principal del SCWRL4 es un método exacto, pero en algunos casos el método no converge y es cuando la heurística entra en juego. El SCWRL4 búsca el óptimo global pero tiene un tiempo máximo para

encontrarlo, y si transcurre dicho tiempo y no se encuentra la solución, éste devuelve una aproximación, de calidad no garantizada. El SCWRL4 utiliza su propia biblioteca de rotámeros, la cual no está disponible aún (Krivov *et al.*, 2009).

El método OPUS-Rota se basa en recocido simulado, por lo tanto no garantiza encontrar el óptimo global. Su función objetivo, incorpora término únicos (Lu *et al.*, 2008a), y se demostró experimentalmente que cada término produce una mejora en la calidad de las soluciones; además utiliza una biblioteca de rotámeros dependiente (Dunbrack y Karplus, 1993).

El SCWRL4 es el método más reciente, y sólo se presentó la comparación con su versión previa (SCWRL3) (Krivov *et al.*, 2009). El método OPUS-Rota se comparó con el SCWRL3, y otras heurísticas, en la cual, el OPUS-Rota tuvo mejores resultados (Lu *et al.*, 2008b).

Aunque en estos trabajos se presentan las medidas de exactitud absoluta, estos valores no se pueden comparar ya que los resultados se generaron con conjuntos de prueba diferentes. Además, los resultados presentados en el método SCWRL4 filtran los residuos por su densidad de electrones (Shapovalov y Dunbrack, 2007), medida que es de reciente creación, y que el OPUS-Rota no toma en cuenta.

El objetivo entonces, es comparar los métodos SCWRL4 y OPUS-Rota de una manera justa, es decir, encontrar un conjunto de pruebas que sea heterogéneo y representativo para realizar esta comparación, además de tomar los mismos criterios al momento de presentar los resultados.

IV.2.1 Características del conjunto de pruebas

La selección de un conjunto de casos de prueba es un paso fundamental en la comparación de métodos para este problema. Uno de los aportes principales de este trabajo es la definción de dicho conjunto, ya que no se cuenta con un cojunto de pruebas que sea heterogéneo y que se pueda utilizar para hacer comparaciones entre métodos de predicción de estructura.

Para seleccionar el conjunto de pruebas se definen diferentes criterios, esto para asegurar la buena calidad de los modelos utilizados. Se seleccionaron proteínas del PDB con las siguientes características.

- Un sólo dominio bajo la clasificación SCOP.

- Resolución igual o menor a 2 Å.
- Factor R igual o menor a 0.2.
- Una sola cadena.
- 40 a 400 aminoácidos.
- Clasificación SCOP dentro de la clase a, b, c o d.
- Identidad máxima de secuencia del 25%.
- Método experimental a través de cristalografía por rayos X.

Para filtrar por resolución y factor R se utilizó el programa S2C (Wang y Dunbrack, 2010). Para filtrar por número de cadenas y cantidad de aminoácidos se utilizó la información de los archivos PDB. Para filtrar por método experimental y por identidad de secuencia se utilizó el servidor PISCES (Wang y Dunbrack, 2003).

Los filtros de SCOP se seleccionaron, ya que se quiere saber si la clasificación SCOP es un determinante para la calidad de las soluciones, mientras que los demás filtros se utilizan para asegurar la calidad de los modelos seleccionados para el conjunto de pruebas.

En la Tabla X se muestran los identificadores de las proteínas del PDB que forman parte del conjunto de pruebas. En total son 770 entradas las que constituyen el conjunto de pruebas con las características de calidad mencionadas anteriormente. Recuerde que en el PDB existen aproximadamente 70 mil estructuras.

IV.2.2 Experimentos

Para realizar los experimentos necesarios para comparar los algoritmos SCWRL4 y OPUS-Rota se utilizaron los programas ejecutables[1] que se encuentran disponibles para fines académicos.

[1]Los programas ejecutables se consiguen bajo licencias para uso académico en los siguientes sitios web.
SCWRL4: http://dunbrack.fccc.edu/scwrl4/
OPUS-Rota: http://sigler.bioch.bcm.tmc.edu/MaLab/soft/opus_rota.html

Tabla X. Conjunto de pruebas.

1A2J, 1A3C, 1A53, 1A68, 1A6M, 1ABA, 1AKO, 1AMF, 1AMX, 1ARB, 1B0U, 1B2V, 1B68, 1BAM, 1BDO, 1BEA, 1BGF, 1BJ7, 1BKF, 1BKR, 1BM8, 1BRT,
1BSO, 1BUD, 1BXE, 1BY2, 1BYI, 1BYR, 1C3J, 1C3P, 1C75, 1C7K, 1CC8, 1CEI, 1CEO, 1CHD, 1CNV, 1CPN, 1CPQ, 1CQY, 1CUJ, 1CV8, 1CXQ, 1CY5,
1CZS, 1D2N, 1D3H, 1D4O, 1DCS, 1DF7, 1DG6, 1DG9, 1DHN, 1DK8, 1DLW, 1DQG, 1DQY, 1DUN, 1DUS, 1DXJ, 1DYP, 1E58, 1EDG, 1EJO, 1ELJ, 1EOK,
1EP0, 1ES9, 1EUR, 1EUW, 1EVF, 1EW4, 1EYE, 1EZM, 1EZW, 1F1E, 1F2J, 1F7L, 1FAZ, 1FCQ, 1FCY, 1FK5, 1FNA, 1FO8, 1FRW, 1FUA, 1FX2, 1FYE,
1G2R, 1G5T, 1G66, 1GA8, 1GBS, 1GCI, 1GEW, 1GMX, 1GNY, 1GP6, 1GPP, 1GPR, 1GQ8, 1GQV, 1GR3, 1GSI, 1GSJ, 1GUI, 1GV9, 1GVD, 1GVG, 1GW1,
1GWM, 1GWU, 1GXN, 1GXQ, 1GXU, 1GYN, 1GZ8, 1HOA, 1HOS, 1H1D, 1H6L, 1H7O, 1H7L, 1HBK, 1HCV, 1HD2, 1HDO, 1HH8, 1HQO, 1HUF, 1HUW, 1HV6,
1HXI, 1HXN, 1HZT, 1IOV, 1I27, 1I2T, 1I4O, 1I6P, 1I8O, 1I9S, 1IAB, 1ICX, 1IFC, 1IJB, 1ILK, 1IM5, 1IOM, 1IQQ, 1IQZ, 1IUL, 1IUZ, 1IX4,
1IXH, 1IZC, 1JOP, 1J27, 1J5X, 1J8Q, 1J8U, 1J98, 1J9B, 1JB3, 1JBE, 1JEO, 1JER, 1JG1, 1JHG, 1JJF, 1JK7, 1JL6, 1JM1, 1JNI, 1JPC, 1JYH,
1K12, 1K1B, 1K3O, 1K51, 1K5C, 1K77, 1K7C, 1K8U, 1KCQ, 1KFR, 1KGD, 1KLL, 1KLX, 1KM4, 1KNB, 1KNM, 1KOE, 1KP6, 1KR7, 1KT6, 1KUX, 1KWF,
1KYP, 1L3P, 1L6P, 1L9L, 1LB3, 1LJO, 1LKE, 1LKI, 1LL2, 1LMI, 1LN4, 1LO7, 1LRO, 1LRI, 1LRK, 1LST, 1LU4, 1LV7, 1LYV, 1LZJ, 1LZK, 1M1Q,
1M2J, 1M8Z, 1MC2, 1MDC, 1ME3, 1MFM, 1MG4, 1MHN, 1MJ4, 1MJ5, 1MJC, 1MSC, 1MSK, 1MUG, 1MUN, 1MUW, 1MVL, 1N1F, 1N3L, 1N4O, 1N55, 1N8N,
1N8U, 1N93, 1NAR, 1NB9, 1NC5, 1NEP, 1NF9, 1NFP, 1NKD, 1NLS, 1NNF, 1NNH, 1NNX, 1NOG, 1NOX, 1NPK, 1NPU, 1NRG, 1NSJ, 1NTE, 1NWA, 1NWZ,
1NZJ, 1O08, 1O1X, 1O1Y, 1O1Z, 1O22, 1O3U, 1O4R, 1O4V, 1O4W, 1O4Y, 1O50, 1O8X, 1O9G, 1OAA, 1OAP, 1OD3, 1OD6, 1ODM, 1OEJ, 1OEW, 1OHO,
1OJQ, 1OJR, 1OKO, 1OKS, 1OOT, 1OPD, 1OS6, 1OZN, 1P3C, 1P4C, 1P5F, 1P5X, 1P5Z, 1P9O, 1P99, 1PA7, 1PB7, 1PBJ, 1PDO, 1PHP, 1PJX, 1PMH,
1POC, 1PQE, 1PUC, 1PWA, 1PZ4, 1PZT, 1QON, 1QOR, 1Q35, 1Q5Z, 1Q8D, 1Q92, 1QCX, 1QDD, 1QJ4, 1QLM, 1QQF, 1QRE, 1QTW, 1QUS, 1QWG, 1QWY,
1QZM, 1ROU, 1R26, 1R29, 1R2Q, 1R3D, 1R5Y, 1R69, 1R8M, 1REC, 1RH9, 1RI6, 1RIS, 1RLH, 1RLJ, 1RQW, 1RTQ, 1RU4, 1RW1, 1RWJ, 1RWR, 1RXE,
1S2W, 1S7Z, 1S9U, 1SBP, 1SBX, 1SDI, 1SFP, 1SGW, 1SHU, 1SK4, 1SN7, 1SNC, 1SQ9, 1SRA, 1SRV, 1SUR, 1SUU, 1SVI, 1SVK, 1SWX, 1SX7, 1SYY,
1TO7, 1T2I, 1T3Y, 1T46, 1T6E, 1T8K, 1TBF, 1TCA, 1TD4, 1THF, 1TIB, 1TIF, 1TJX, 1TJY, 1TP6, 1TQG, 1TR9, 1TT8, 1TU9, 1TUH, 1TUK, 1TUW,
1TXJ, 1TZV, 1U14, 1U36, 1UAI, 1UAL, 1UBI, 1UCS, 1UFY, 1UIO, 1UKU, 1UKZ, 1UMG, 1UMH, 1UNQ, 1UOZ, 1UPQ, 1UQ5, 1US5, 1USG, 1UUY, 1UV4,
1UWF, 1UYL, 1V05, 1VOA, 1VOL, 1VHH, 1VJF, 1VJO, 1VK1, 1VK4, 1VKB, 1VKK, 1VL1, 1VLC, 1VLS, 1VMB, 1VMG, 1VMH, 1VP8, 1VPR, 1VQB, 1VR8,
1VYI, 1VYR, 1WON, 1W1G, 1W3L, 1W3U, 1W53, 1W66, 1WBA, 1WC2, 1WD5, 1WDE, 1WHI, 1WHZ, 1WKA, 1WL8, 1WLU, 1WM3, 1WP5, 1WRI, 1WWI, 1WXI,
1WXJ, 1X06, 1X6Z, 1X82, 1X8H, 1X8Q, 1X91, 1XBI, 1XDN, 1XEO, 1XGK, 1XKI, 1XKN, 1XMK, 1XMT, 1XQO, 1XTE, 1XTP, 1XVO, 1YOK, 1Y8A, 1Y93,
1YE8, 1YFQ, 1YHT, 1YMK, 1YOY, 1YPC, 1YU5, 1YWF, 1YZM, 1Z4R, 1Z67, 1Z6M, 1Z6N, 1Z7O, 1ZCE, 1ZGK, 1ZK4, 1ZMA, 1ZND, 1ZRN, 1ZUU, 1ZV9,
1ZZK, 1ZZM, 2A14, 2A15, 2A4V, 2A6Z, 2A7B, 2A84, 2ABK, 2ABS, 2AH5, 2AI2, 2ALI, 2AM9, 2AMH, 2AQ8, 2ASF, 2AZW, 2B06, 2BOA, 2B3M, 2B4W,
2B61, 2B65, 2B69, 2B8M, 2BBH, 2BKF, 2BRF, 2BWQ, 2BZ1, 2C0H, 2C3F, 2C60, 2C71, 2C9Q, 2CCB, 2CCQ, 2CCV, 2CCW, 2CFE, 2CHH, 2CIC, 2CJ7,
2CNQ, 2CTC, 2CU9, 2CUL, 2CW4, 2CWY, 2CXH, 2CXV, 2CYG, 2D28, 2D2J, 2D4P, 2D81, 2DFB, 2DJH, 2DRI, 2DUY, 2E3B, 2E3H, 2EBN, 2ECQ, 2END,
2ENG, 2ERF, 2ERL, 2ESK, 2ET1, 2EUT, 2EWO, 2EWH, 2EWR, 2F1N, 2F71, 2F9F, 2FBH, 2FDN, 2FDR, 2FE5, 2FFC, 2FG1, 2FI1, 2FK8, 2FOU, 2FSQ,
2FSR, 2FSU, 2FUF, 2FUP, 2FUZ, 2FVV, 2FVY, 2FZP, 2G2C, 2G3A, 2G4O, 2G62, 2G64, 2G7O, 2G9F, 2GDM, 2GFO, 2GGC, 2GHS, 2GKP, 2GM6, 2GNP,
2GPI, 2GS5, 2GUI, 2GUX, 2GVK, 2H5C, 2HCF, 2HEW, 2HJE, 2HK6, 2HKV, 2HLJ, 2HLY, 2HNG, 2HTS, 2HUH, 2HVF, 2HXO, 2HXM, 2I1B, 2I5U, 2I6C,
2I9C, 2IA7, 2IAY, 2ICA, 2ICG, 2IE7, 2IGD, 2IHK, 2IM9, 2IMQ, 2INB, 2ISB, 2IUW, 2IVY, 2IW1, 2IYV, 2IZK, 2J6B, 2J8K, 2JDC, 2JEK, 2LIS,
2MCM, 2MHR, 2NLR, 2NML, 2NN8, 2NQW, 2NR7, 2NRK, 2NSZ, 2NW6, 2NWV, 2NX2, 2NXF, 2O0M, 2O1A, 2O2X, 2O8P, 2O9U, 2OB5, 2OC5, 2OD5, 2OEB,
2OH3, 2ONS, 2OO3, 2OQZ, 2OSO, 2OT9, 2OU6, 2OVO, 2P3K, 2P4O, 2P5K, 2PB1, 2PII, 2PMR, 2PNW, 2PPQ, 2PPX, 2PQ7, 2PST, 2PTD, 2PTH, 2PVB,
2PXR, 2Q3M, 2Q3P, 2Q3T, 2Q3W, 2Q4M, 2Q4N, 2Q7W, 2QED, 2QIA, 2QSB, 2QSW, 2QWC, 2QYQ, 2R2Z, 2R31, 2R4Q, 2R9F, 2RBK, 2RFR, 2RH2, 2RH3,
2RHW, 2RIU, 2RN2, 2SAK, 2UYQ, 2V3G, 2V3K, 2V7F, 2VB1, 2VBU, 2VEP, 2VMH, 2VPA, 2VRY, 2YVT, 2YXF, 2Z3V, 2Z94, 2Z98, 2ZFI, 3B7C, 3BCJ,
3BFP, 3BI7, 3BOE, 3BZT, 3CKM, 3DM8, 3DNI, 3DUE, 3E99, 3EBT, 3EBY, 3EJV, 3ELN, 3EN8, 3IL8, 3NUL, 3SIL, 3VUB, 4RHN, 5CSM, 6FIV, 8ABP

Ambos métodos toman como entrada archivos PDB donde se indica la secuencia de aminoácidos de una proteína específica, junto con las coordenadas de los átomos de la columna vertebral; y generan como salida archivos PDB con las coordenadas de todos los átomos (cadena lateral y columna vertebral), donde las coordenadas de los átomos de la columna vertebral son las mismas que las de la entrada, y las coordenadas de los átomos de la cadena lateral son las predicciones hechas por los métodos.

Debido a que el SCWRL4 es un método determinístico, éste se ejecutó una vez por cada entrada del conjunto de pruebas.

El OPUS-Rota es un método basado en recocido simulado, por lo tanto, para realizar un análisis es necesario ejecutarlo varias veces para una misma entrada, ya que los métodos basados en recocido simulado son métodos aleatorios y para hacer un análisis estadístico se requiere no una, sino varias ejecuciones del mismo método, por lo que se requirió que generara 30 archivos de salida por cada elemento del conjunto de pruebas.

IV.3 Resultados

Una vez que se obtuvieron los 770 archivos PDB de salida del SCWRL4, y los 23100 ($770 \times 30 = 23100$) archivos PDB de salida del OPUS-Rota se puede hacer un análisis de la calidad de las soluciones de ambos métodos. En esta sección se presentan los resultados obtenidos al analizar tanto el conjunto de pruebas como las predicciones de los métodos seleccionados.

Primero se presentan los análisis realizados al conjunto de pruebas para mostrar que los modelos utilizados son de alta calidad, y son representativos para poder realizar la comparación entre los métodos.

Posteriormente se describe el análisis realizado sobre las predicciones de ambos métodos para determinar si existe alguna característica que haga que un método genere mejores predicciones que el otro.

IV.3.1 Conjunto de pruebas

Longitud de secuencias

En la Figura 13 se puede observar la distribución de la longitud de secuencias en tres conjuntos: el conjunto de pruebas (13(a)), el conjunto de secuencias de UniProtKB/-

Swiss-Prot (13(b)) y el conjunto de secuencias de UniProtKB/TrEMBL (13(c)).

UniProtKB/Swiss-Prot (uni, 2010a) y UniProtKB/TrEMBL (uni, 2010b) son bases de datos de secuencias de proteínas que contienen más de 500 mil y más de 11 millones de secuencias, respectivamente. La diferencia entre ambas bases de datos es que el proceso de anotación, en UniProtKB/Swiss-Prot es manual, mientras que en UniProtKB/TrEMBL es automático (Wu *et al.*, 2006; Boeckmann *et al.*, 2003).

Los promedios de longitud de secuencia para UniProtKB/Swiss-Prot y UniProtKB/TrEMBL son 352 y 321 aminoácidos por secuencia, respectivamente. La longitud de las secuencias no sigue una distribución normal y está sesgada a la izquierda. Debido a que nuestro conjunto de pruebas sólo contempla secuencias con longitud entre 40 y 400 aminoácidos, su distribución es ligeramente diferente a las otras dos.

Resolución

En la Sección II.1.6 se describe la importancia de este parámetro para evaluar la calidad de una estructura. En la Figura 14 se muestra la distribución de los casos de prueba por resolución. En esta figura se puede observar que todos los casos tienen resolución menor o igual a 2 Å, lo cual es un buen indicador de la calidad de los modelos utilizados en el conjunto de pruebas.

Factor-R

En la Sección II.1.6 se describe el factor-R y su importancia para asegurar la calidad de una estructura. En la Figura 15 se muestra la distribución del factor R. En la Tabla XI se muestran cuatro casos para los cuales el programa PISCES (utilizado para hacer la Figura 15) reportó un factor R mayor a 0.2, junto con los valores reportados por el sitio web del PDB y el programa S2C. Se puede ver que las diferencias no son grandes, y además que el programa S2C se acerca más a lo reportado en el PDB, por lo que es mejor utilizar el S2C que el del programa PISCES para filtrar por el factor R, tal como se hizo para generar el conjunto de pruebas.

El problema de tener datos discrepantes se presenta comúnmente en bioinformática, es uno de los problemas que hacen aún más difícil el análisis de los datos, ya que hay que ser muy cuidadosos al utilizar datos provenientes de distintas fuentes.

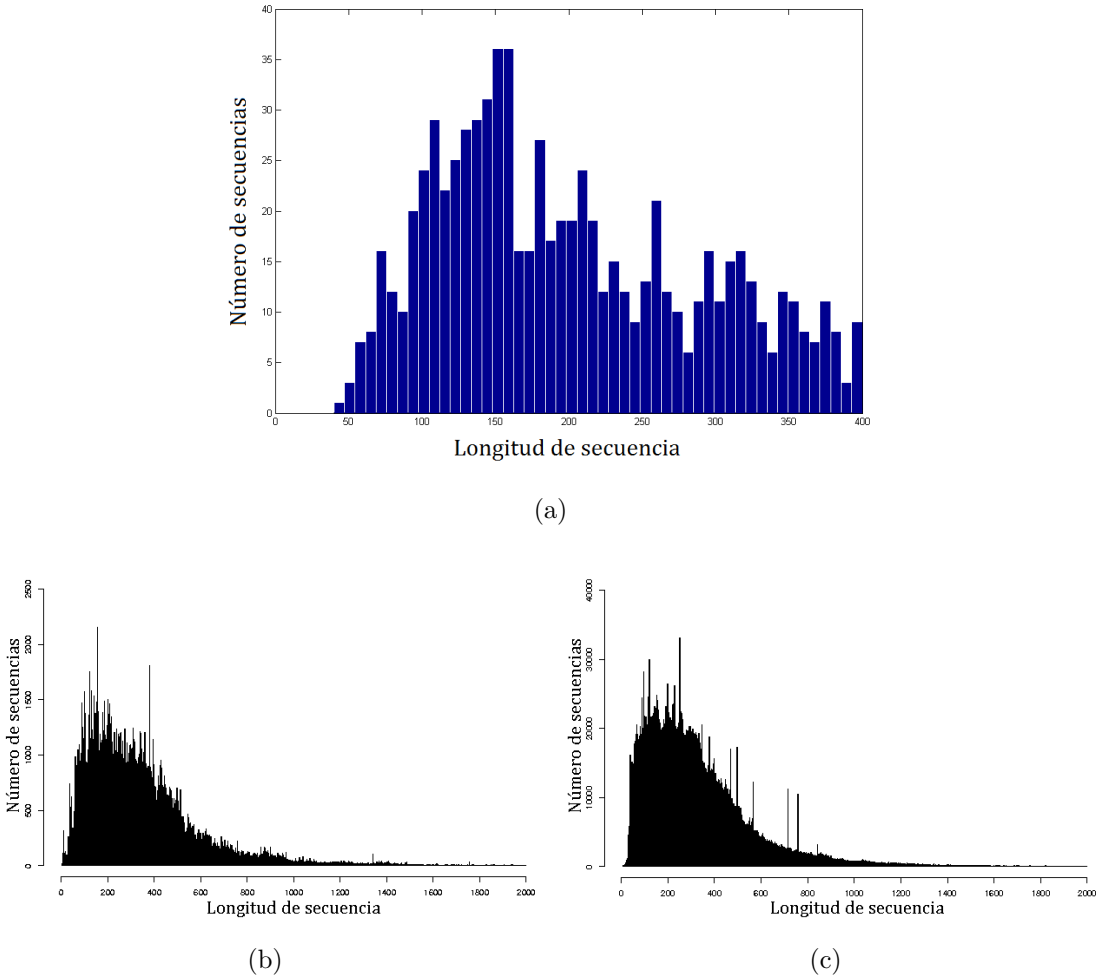


Figura 13. Distribución de longitud de secuencias. (a) Conjunto de pruebas. (b) UniProtKB/Swiss-Prot. (c) UniProtKB/TrEMBL.

Tabla XI. Factor R.

PDB ID	PDB	PISCES	S2C
1HUF	0.192	0.22	0.1921
1L3P	0.197	0.21	0.1970
1NPU	0.198	0.21	0.1980
1YZM	0.196	0.21	0.1956

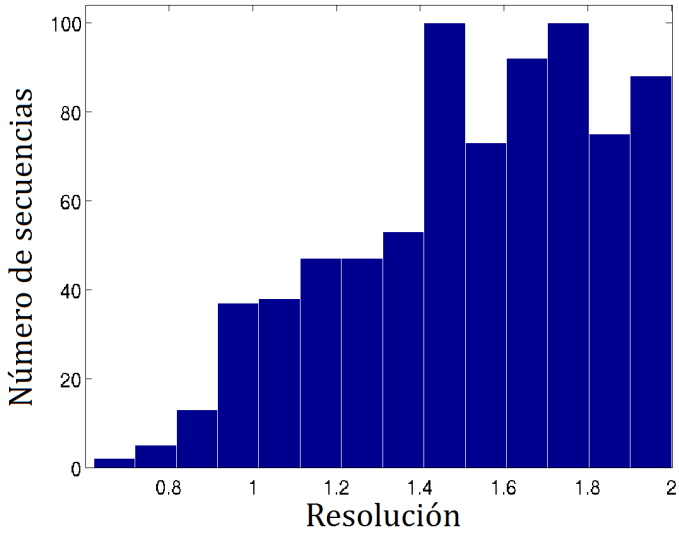


Figura 14. Distribución de casos por resolución.

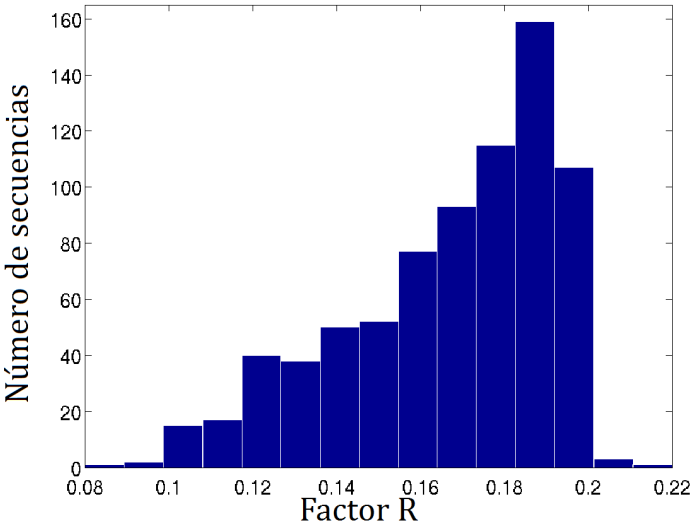


Figura 15. Distribución de casos por factor R.

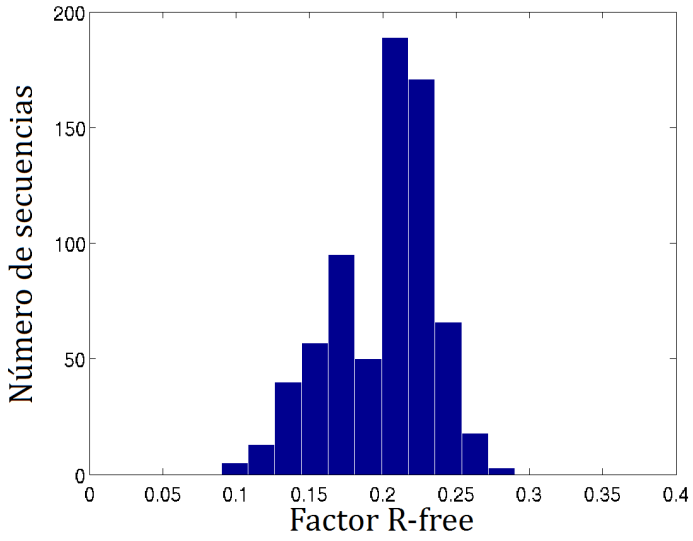
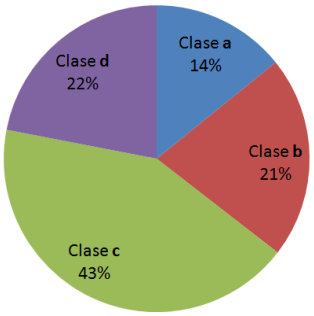


Figura 16. Distribución de casos por factor R-free.

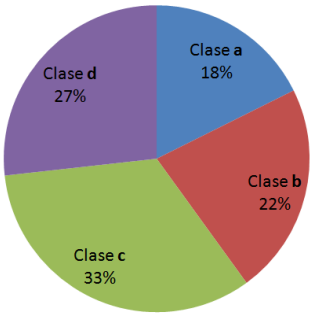
Factor R-Free

El factor R-Free es parecido al factor-R. Se obtiene para medir que tan parecido es un modelo o estructura generada con los datos experimentales. Si el modelo encajara perferctamemte a los datos experimentales, estos factores son 0; sin embargo, en la práctica esto no es así. El factor-R se obtiene en el proceso de refinamiento, al ajustar el modelo a los datos experimentales; sin embargo, al utilizar este valor como objetivo en el proceso de optimización, éste está sesgado. Una manera de reducir el sesgo es utilizar sólo una fracción de los datos en el proceso de refinamiento (normalmente se utiliza el 90%) y el resto dejarlos sin modificación, y utilizarlos para obtener el factor R-free al final del proceso de refinamiento.

En la Figura 16 se muestra la distribución del factor R-free para el conjunto de pruebas. Para que un modelo sea de buena calidad también se puede usar este valor como referencia, tomando un valor de corte para el factor R-free de 0.4 es suficiente (Gu y Bourne, 2009), y se ve que para el conjunto de prueba, las secuencias con mayor R-free reportado es 0.29, por lo que se puede decir que se están usando modelos de alta calidad. Sin embargo, hay 63 casos para los cuales no se reportó el factor R-free y no se incluyen en el histograma. Por esta razón, es mejor filtrar por el factor R, ya que es más común encontrar modelos para los cuales no se registró el factor R-free que el factor R.



(a) Residuos agrupados por clase SCOP.



(b) Casos agrupados por clase SCOP.

Figura 17. Distribución del conjunto de pruebas según las clases del SCOP.

SCOP

En las Figuras 17(a) y 17(a) se muestran los porcentajes de residuos y casos del conjunto de pruebas, respectivamente, agrupados por la clase de SCOP (Murzin *et al.*, 1995) a la que pertenecen. Las clases de SCOP que se incluyeron en el conjunto de pruebas son las siguientes.

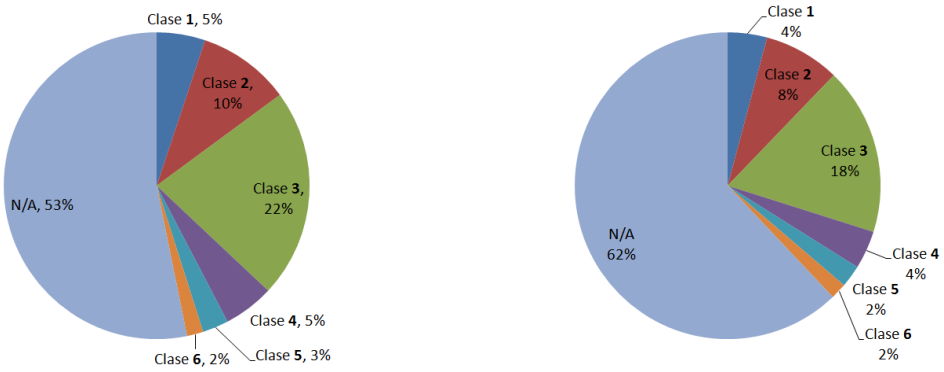
- **Clase a**, proteínas cuyas estructuras secundarias predominantes son hélices-α.
- **Clase b**, proteínas cuyas estructuras secundarias predominantes son hojas-β.
- **Clase c** proteínas con hélices-α y hojas-β (no separables).
- **Clase d**, proteínas con hélices-α y hojas-β (separables).

Se puede observar que la clase predominante es la clase c, pero de las otras clases hay suficientes residuos y casos para poder hacer un análisis por clase.

EC

El número EC es una clasificación jerárquica de enzimas, según las reacciones químicas que catalizan. Por ejemplo, la proteína con PDB ID=1JVA es una endonucleasa con número EC 3.6.1.34. Dónde el número 3 representa el grupo de las hidrolasas y el resto es la subclase (Moss, 2010).

En las Figuras 18(a) y 18(b) se muestran los porcentajes de residuos y casos, respectivamente, agrupados por grupo EC. Se puede observar que el 53% de los residuos



(a) Residuos agrupados por clase EC. (b) Casos agrupados por clase EC.

Figura 18. Distribución del conjunto de pruebas según las clases EC. Las clases EC son: (1) oxidoreductasas, (2) transferasas, (3) hidrolasas, (4) liasas, (5) isomerasas y (6) ligasas.

(77428 residuos de 145572), que pertenecen al 62% de los casos (478 casos de 770) no pertenecen a alguna clase EC. Esto se debe a que la clasificación EC es para enzimas, por lo que el 62% de los casos del conjunto de pruebas no son enzimas, o bien, aun no se han clasificado como tales.

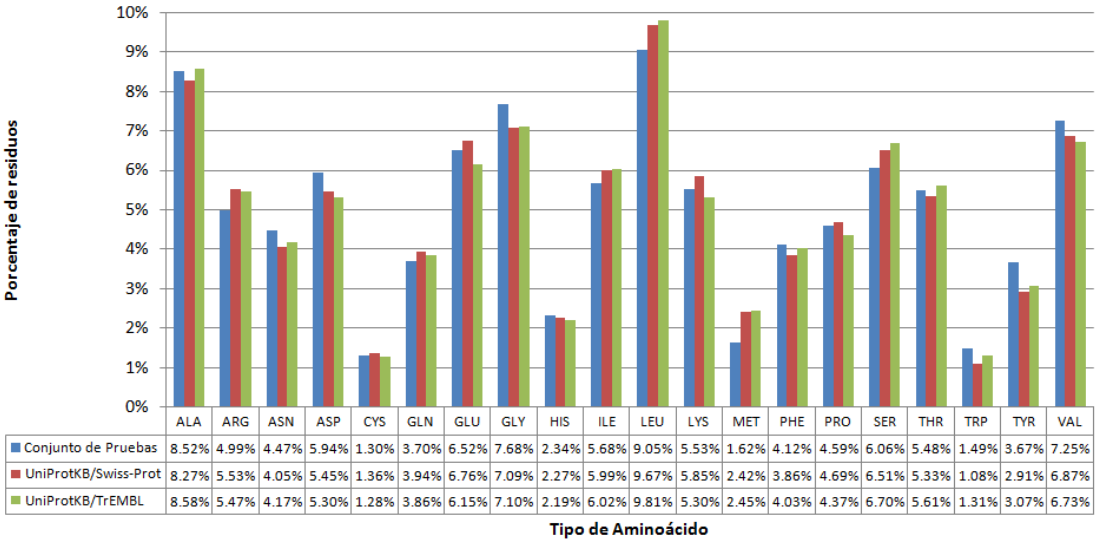
Distribución de residuos por tipo de aminoácido

En la Figura 19 se muestran las distribuciones de los residuos por tipo de aminoácido tanto para el conjunto de pruebas como para las secuencias anotadas en UniProtKB/-Swiss-Prot y UniProtKB/TrEMBL. Se puede ver que la distribución para el conjunto de pruebas es similar al de UniProtKB/Swiss-Prot y al de UniProtKB/TrEMBL, lo que indica que nuestro conjunto sigue un patrón similar en cuanto a tipo de aminoácidos que aquel que siguen las proteínas anotadas hasta el momento en ambas bases de datos.

IV.3.2 SCWRL4 vs. OPUS-Rota

Calidad global

En la Figura 20 se puede observar la calidad de las soluciones de ambos métodos utilizando diferentes umbrales para indicar si un residuo fue correctamente predicho o no. En la Tabla XII se muestran los valores obtenidos para cada medida. Los porcentajes se obtuvieron al dividir la cantidad de residuos totales cuyos ángulos en cuestión están dentro del umbral, entre el total de residuos que deben considerarse para



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Conjunto de Pruebas	8.52%	4.99%	4.47%	5.94%	1.30%	3.70%	6.52%	7.68%	2.34%	5.68%	9.05%	5.53%	1.62%	4.12%	4.59%	6.06%	5.48%	1.49%	3.67%	7.25%
UniProtKB/Swiss-Prot	8.27%	5.53%	4.05%	5.45%	1.36%	3.94%	6.76%	7.09%	2.27%	5.99%	9.67%	5.85%	2.42%	3.86%	4.69%	6.51%	5.33%	1.08%	2.91%	6.87%
UniProtKB/TrEMBL	8.58%	5.47%	4.17%	5.30%	1.28%	3.86%	6.15%	7.10%	2.19%	6.02%	9.81%	5.30%	2.45%	4.03%	4.37%	6.70%	5.61%	1.31%	3.07%	6.73%

Figura 19. Distribución de residuos por tipo de aminoácido.

Tabla XII. Exactitud abslouta.

Umbral	SCWRL4				OPUS-Rota			
	$\chi_1(\%)$	$\chi_{1,2}(\%)$	$\chi_{1,2,3}(\%)$	$\chi_{1,2,3,4}(\%)$	$\chi_1(\%)$	$\chi_{1,2}(\%)$	$\chi_{1,2,3}(\%)$	$\chi_{1,2,3,4}(\%)$
10	69.24	41.10	15.63	11.34	69.22	40.53	14.80	10.74
20	82.56	61.24	31.64	24.80	82.75	60.89	30.95	24.00
30	84.78	66.35	37.90	28.91	85.08	66.47	37.49	28.37
40	85.62	68.31	41.23	30.97	85.99	68.66	41.13	30.23

la medida. Con el objeto de dejar claro el procedimiento para obtener estas medidas, se propone el siguiente ejemplo.

Del total de las 770 secuencias, se tienen 153850 aminoácidos, de los cuales 145572 tienen todos los átomos principales y son los que se utilizan para reportar las diferentes medidas de calidad. Sin embargo, para el método OPUS-Rota se generaron 30 predicciones por secuencia, entonces en ese caso se tienen 4367160 aminoácidos a analizar.

Para el ejemplo de la medida $\chi_{1,2,3,4}(\%)$ del método SCWRL4 se tienen en total 15320 aminoácidos a considerar, es decir, de los 145572 residuos totales, sólo 15320 son del tipo arginina (ARG) o lisina (LYS), que son los aminoácidos que tienen hasta el ángulo χ_4. De esos 15320 aminoácidos, sólo 1738 tienen una diferencia menor o igual a 10^o en todos los ángulos, desde χ_1 hasta χ_4, respecto a los ángulos originales del PDB, dando como resultado 11.34% para la medida $\chi_{1,2,3,4}(\%)$ con umbral de 10^o del método

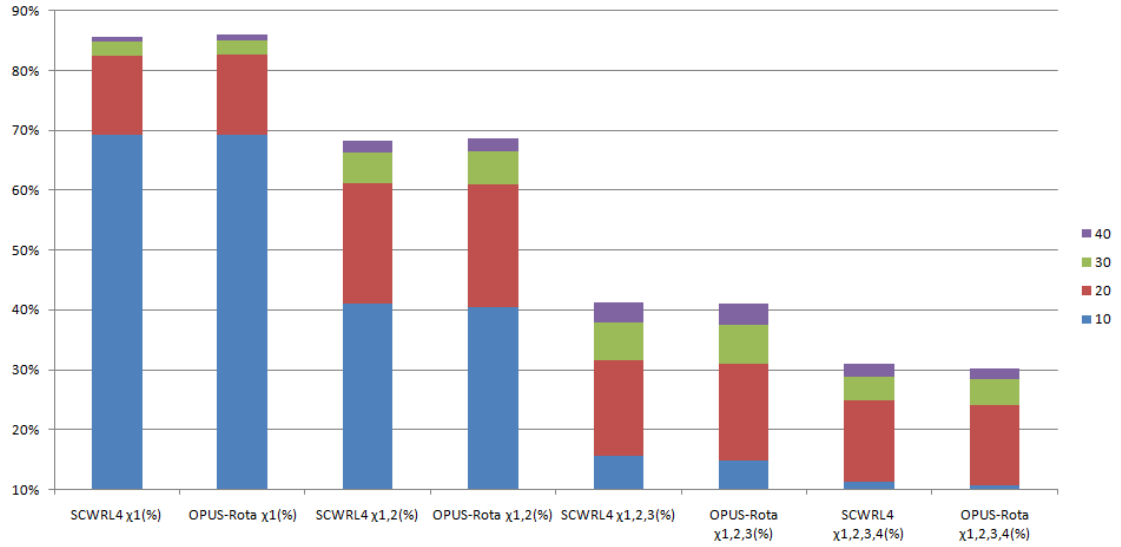


Figura 20. Exactitud absoluta.

SCWRL4.

Por el contrario, para el método OPUS-Rota, se tiene un total de 459600 (15320 $\times$ 30 = 459600) aminoácidos a considerar, de los cuales, sólo 49339 tienen una diferencia menor o igual a 10^o en todos los ángulos respecto a los águlos originales del PDB, dando como resultado 10.74% para la medida $\chi_{1,2,3,4}(\%)$ con umbral de 10^o del método OPUS-Rota.

Las medidas más comúnmente reportadas por los diferentes métodos son $\chi_1(\%)$ y $\chi_{1,2}(\%)$ con umbral de 40^o, en estos casos, el OPUS-Rota tuvo mejores resultados. Sin embargo, el SCWRL4 es mejor que el OPUS-Rota para las medidas $\chi_{1,2,3}(\%)$ y $\chi_{1,2,3,4}(\%)$.

En la Figuras 21-24 se muestran la distribuciones acumuladas del error de los ángulos de torsión de la cadena lateral predichos (χ_1' - χ_4') respecto a los ángulos originales. Para calcular el error en la predicción del ángulo, utilizamos la siguiente expresión.

$$e(\chi, \chi') = \min(|\chi - \chi'|, 360 - |\chi - \chi'|) \tag{24}$$

De manera similar, en la Figura 25 se muestra la distribución acumulada del RMSD (ver Sección III.3.1) por residuo.

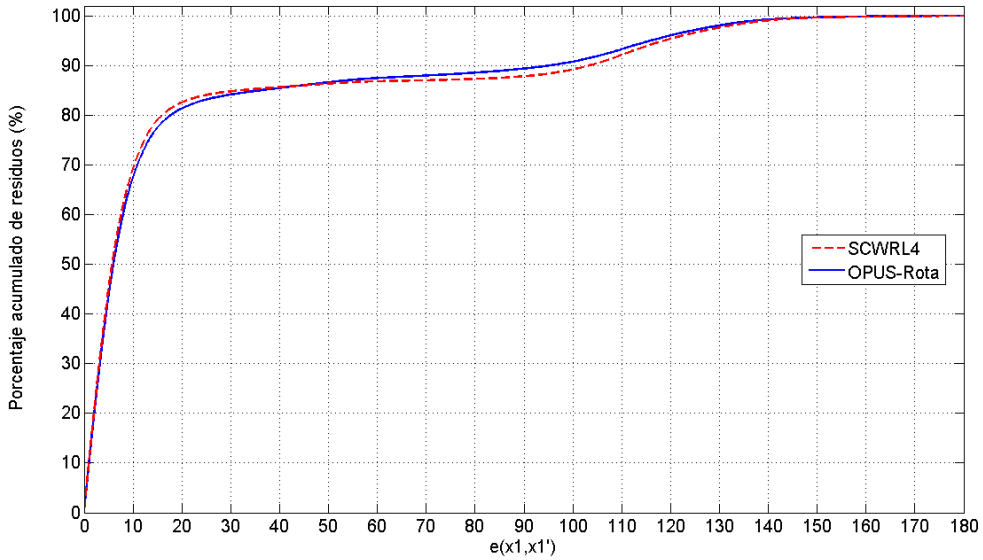


Figura 21. Distribución acumulada del error del ángulo predicho (χ_1') respecto al ángulo χ_1 original.

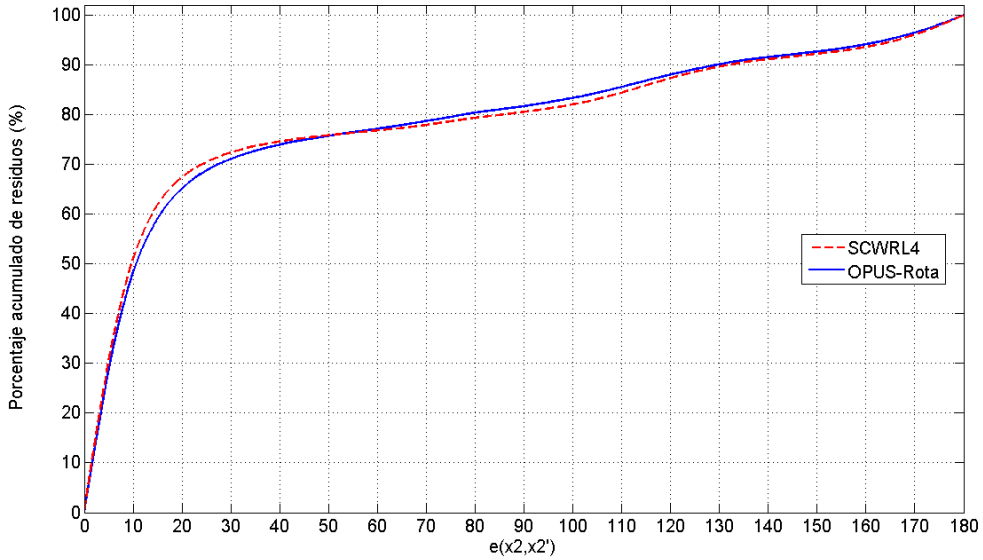


Figura 22. Distribución acumulada del error del ángulo predicho (χ_2') respecto al ángulo χ_2 original.

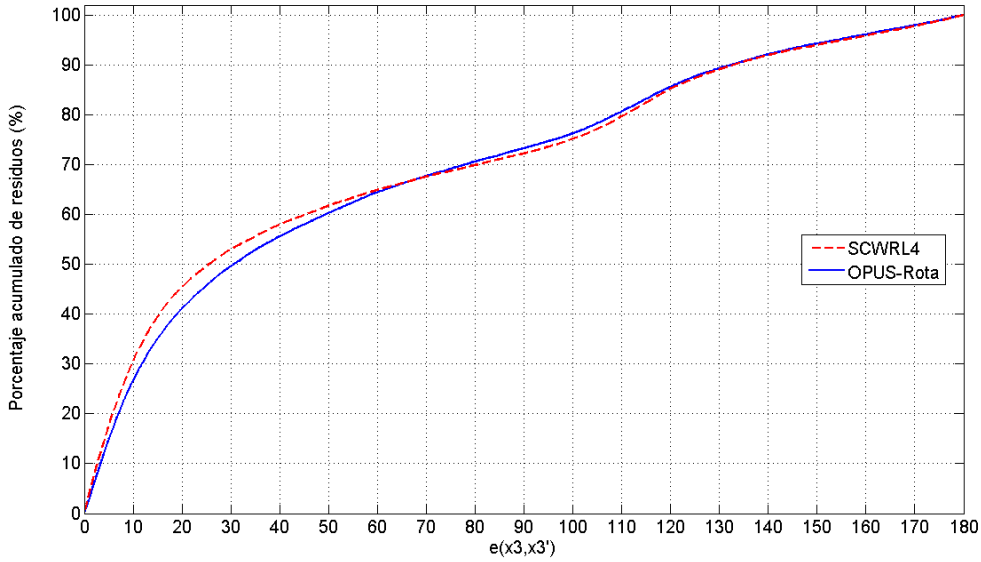


Figura 23. Distribución acumulada del error del ángulo predicho (χ_3') respecto al ángulo χ_3 original.

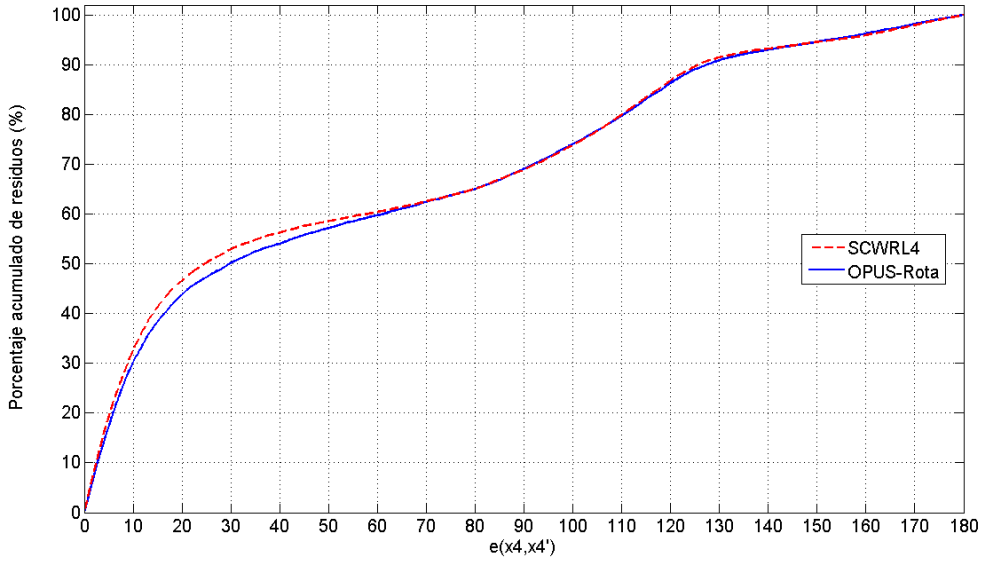


Figura 24. Distribución acumulada del error del ángulo predicho (χ_4') respecto al ángulo χ_4 original.

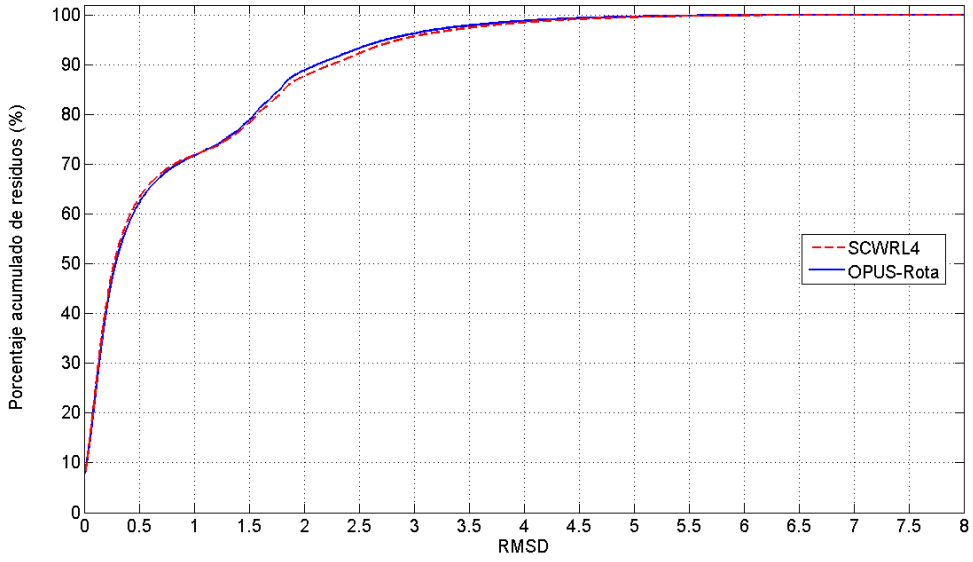


Figura 25. Distribución acumulada del RMSD por residuo.

Calidad por clase del SCOP

El objetivo de este experimento es determinar si la clase SCOP es una determinante para la calidad de las soluciones en alguno de los dos métodos que se están analizando.

En la Figura 26 se muestra la medida $\chi_1(\%)$ agrupada por método y clase del SCOP. Se puede observar que para todas las clases del SCOP, el método OPUS-Rota generó mejores resultados; sin embargo, no hay una diferencia notoria entre la calidad de las soluciones de una clase y otra.

Calidad por clase del EC

El objetivo de este experimento es determinar si la clasificación EC es un factor que interviene para la calidad de las soluciones, es decir, se agrupan las secuencias por clase EC y se determina si algún conjunto sobresale en cuanto a calidad de las predicciones.

En la Figura 27 se muestra la medida $\chi_1(\%)$ para diferentes umbrales, agrupados por clase EC. Se puede observar que para este caso la clase EC no es un factor que marque una diferencia importante en la calidad de las soluciones, pero aún así se puede observar nuevamente la consistencia en la mejora del método OPUS-Rota respecto al método SCWRL4, con excepción de la clase 5.

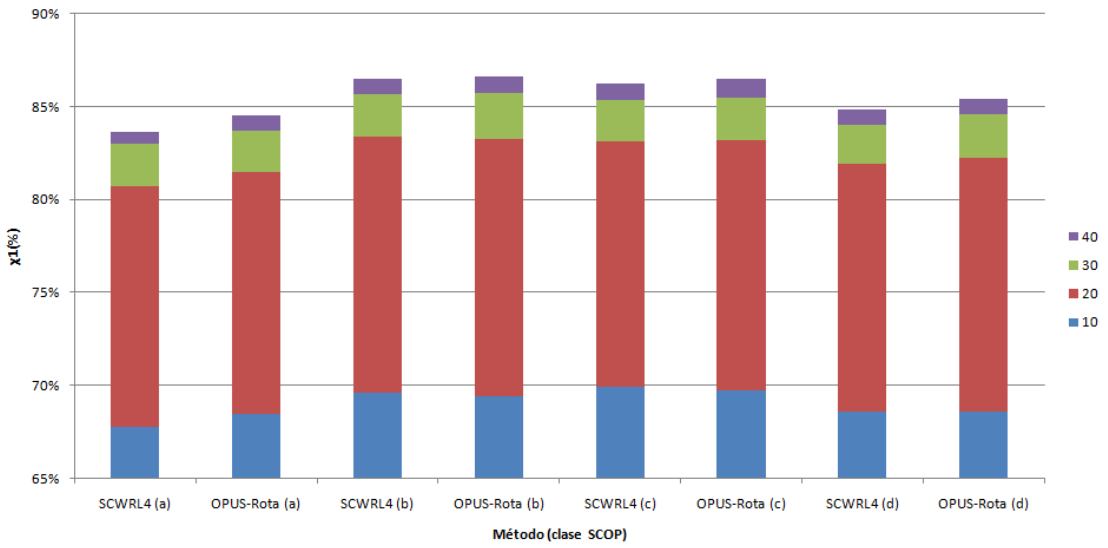


Figura 26. Exactitud Absoluta $\chi_1(\%)$ agrupado por clase del SCOP para diferentes umbrales.

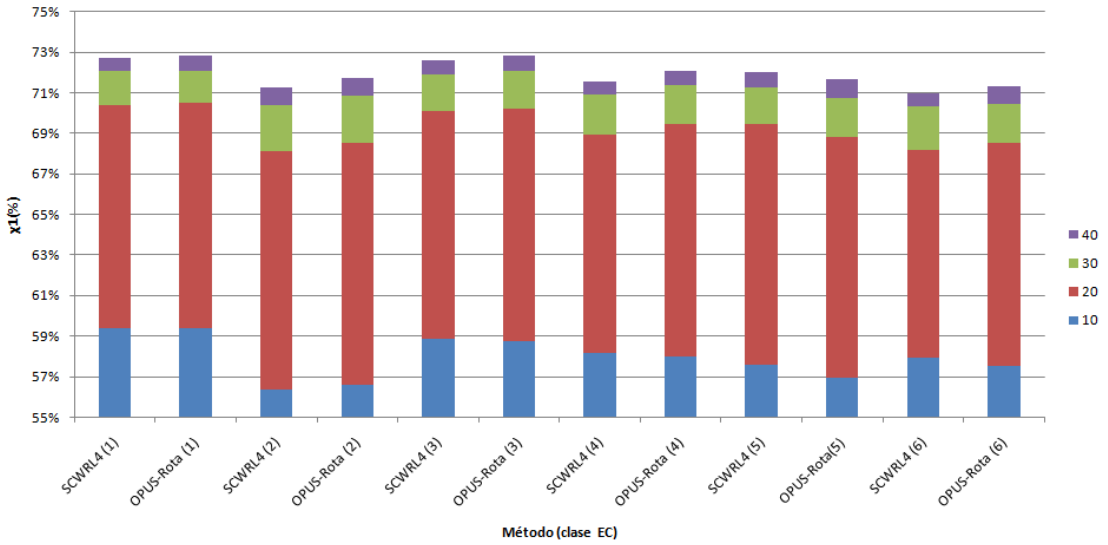


Figura 27. Exactitud absoluta $\chi_1(\%)$ agrupado por clase EC.

Calidad por tipo de aminoácido

El objetivo de este experimento es saber si el tipo de aminoácido es un factor que determina la calidad de las soluciones. El total de residuos se agrupa por tipo y se muestra la medida $\chi_1(\%)$ por grupo. En la Figura 28 se muestran los resultados de este experimento, en el cual se puede observar que el tipo de aminoácido sí afecta la exactitud.

Los aminoácidos isoleucina (ILE), fenilalanina (PHE), tirosina (TYR), valina (VAL), triptófano (TRP) y leucina (LEU) obtuvieron la mejor exactitud para ambos métodos. Como se observa en la Figura 5, estos aminoácidos son no polares (con excepción de la tirosina que se considera anfipático[2]), por lo que se encuentran principalmente en el núcleo de las proteínas, y como se ha mostrado con anterioridad (Lu *et al.*, 2008b), los residuos que se encuentran en el núcleo de las proteínas se pueden predecir con mayor exactitud, mientras que los aminoácidos que son propensos a estar en la superficie, es decir, que son polares, se predicen con menor exactitud.

Esto se puede deber a diferentes causas, una de ellas está relacionada con la flexibilidad que presentan los aminoácidos en la superficie, ya que es sabido que las proteínas se encuentran en constante movimiento, permitiendo que algunos átomos de la superficie cambien su conformación, mientras que los residuos que se encuentran en el centro de las proteínas son más rígidos, dándo estabilidad a la estructura global de las proteínas.

En la Figura 29 se muestra la diferencia de las medidas $\chi_1(\%)$ para los métodos OPUS-Rota y SCWRL4. Si la diferencia es positiva, quiere decir que el OPUS-Rota obtuvo mejores resultados, de lo contrario, el SCWRL4 fue mejor. Se puede observar que para la prolina (PRO) y serina (SER), el SCWRL4 es superior, mientras que para la cisteína (CYS), histidina (HIS), lisina (LYS), metionina (MET), fenilalanina (PHE), tirosina (TYR) y leucina (LEU), el OPUS-Rota es superior.

La prolina tiene una característica que la hace muy diferente al resto de los aminoácidos. Este aminoácido, es el único cuya cadena lateral tienen más de un enlace con la columna vertebral, es decir, además de que el carbono$-\beta$ (CB) esté unido al carbono$-\alpha$ (CA), el carbono$-\delta$ (CD) está unido a través de un enlace covalente al nitrógeno (N) de la columna vertebral. Esto genera mayor rigidez en la cadena lateral, lo cual indica menos conformaciones, y posiblemente sea una cualidad que el método

[2]Anfipático: que contiene a la vez dominios polares y no polares.

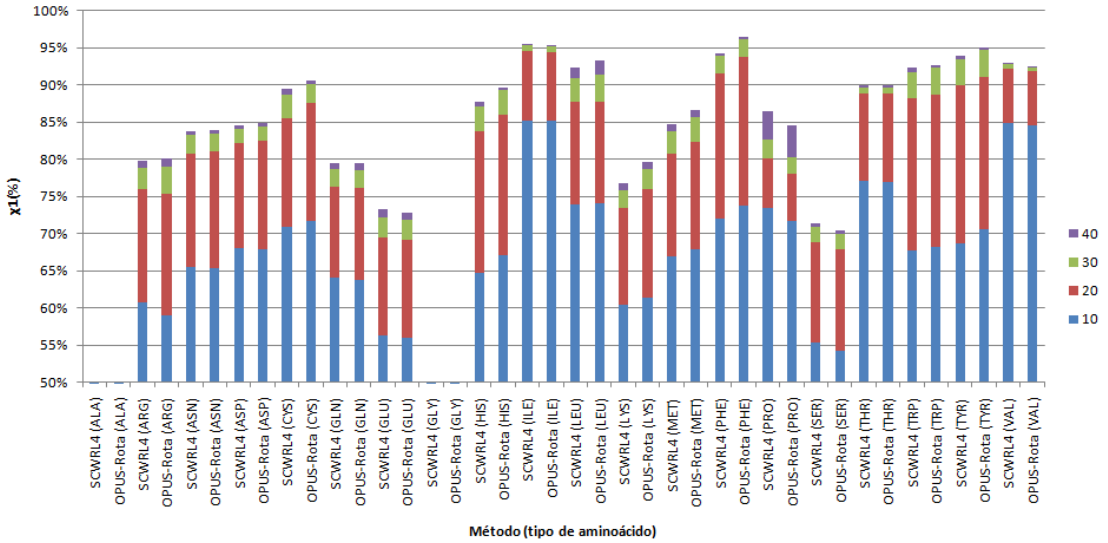


Figura 28. Exactitud absoluta $\chi_1(\%)$ agrupado por tipo de aminoácido para diferentes umbrales.

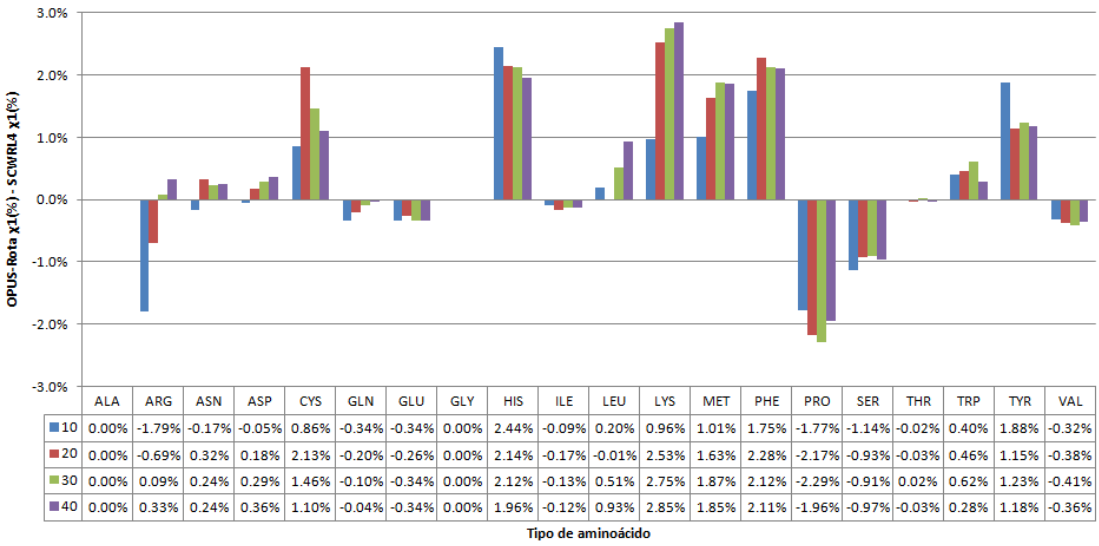
SCWRL4 aprovecha para generar mejores resultados para este aminoácido.

En Krivov *et al.* (2009) se reportan las métricas $\chi_1(\%)$ por tipo de aminoácido, así como de manera global. El estudio lo realizaron sobre un conjunto de 379 proteínas, pero un inconveniente es que para obtener las medidas reportadas sólo utilizaron residuos cuya densidad de electrones esté arriba del percentil 25.

En la Figura 30 se muestran las medidas $\chi_1(\%)$ utlizando únicamente el SCWRL4 para dos conjuntos diferentes. Los conjuntos tienen 379 y 770 proteínas cada uno. El primer conjunto es el conjnto de pruebas de Krivov *et al.* (2009) y el segundo es el nuestro.

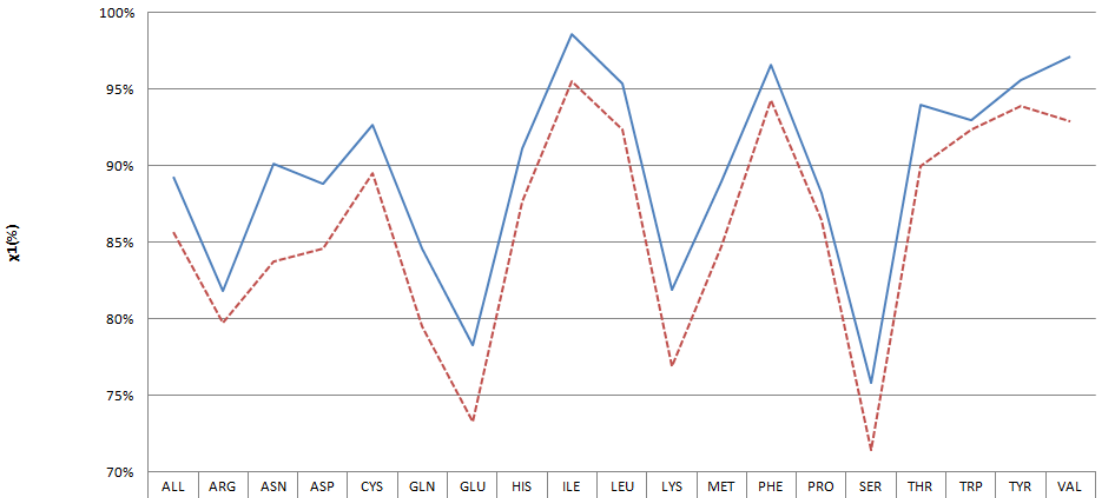
Como se puede observar en la figura, las medidas de ambos conjuntos siguen la misma tendencia; sin embargo, las medidas del conjunto de 770 proteínas (nuestro conjunto de pruebas) reporta menor calidad. Para generar nuestros resultados, se incluyen todos los residuos, y esto genera el desplazamiento hacia abajo de las medidas de nuestro conjunto.

En la Figura 31 se muestra el RMSD promedio por tipo de aminoácido. Se puede observar que para el grupo de las argininas (ARG) se presenta el mayor RMSD promedio; esto se debe a que la forma alargada de la cadena lateral de la arginina genera que la diferencia de los átomos se vaya acumulando, generando un RMSD relativamente grande, comparado a otros residuos con menos átomos en la cadena lateral, u otros



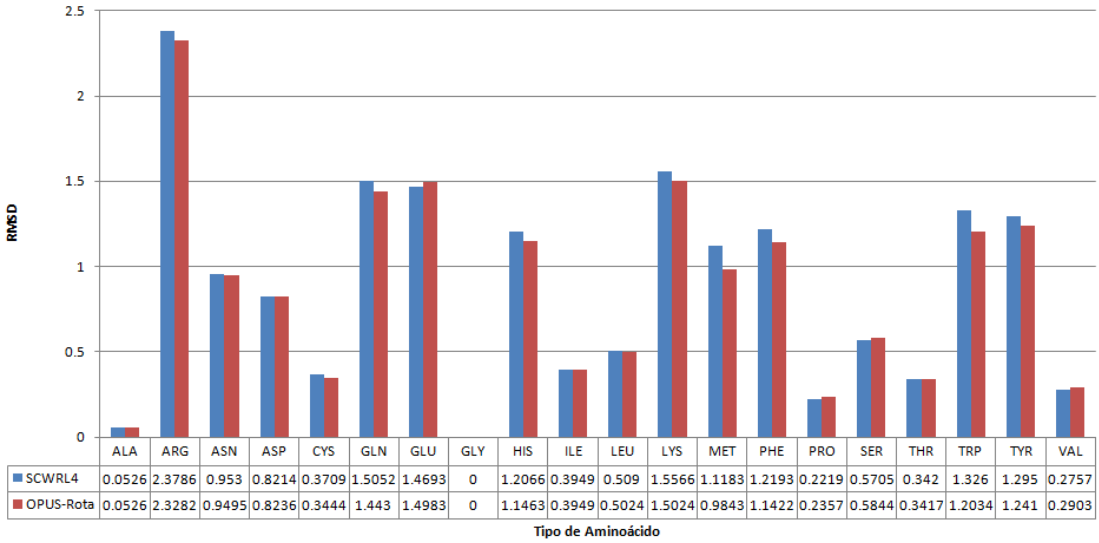
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
10	0.00%	-1.79%	-0.17%	-0.05%	0.86%	-0.34%	-0.34%	0.00%	2.44%	-0.09%	0.20%	0.96%	1.01%	1.75%	-1.77%	-1.14%	-0.02%	0.40%	1.88%	-0.32%
20	0.00%	-0.69%	0.32%	0.18%	2.13%	-0.20%	-0.26%	0.00%	2.14%	-0.17%	-0.01%	2.53%	1.63%	2.28%	-2.17%	-0.93%	-0.03%	0.46%	1.15%	-0.38%
30	0.00%	0.09%	0.24%	0.29%	1.46%	-0.10%	-0.34%	0.00%	2.12%	-0.13%	0.51%	2.75%	1.87%	2.12%	-2.29%	-0.91%	0.02%	0.62%	1.23%	-0.41%
40	0.00%	0.33%	0.24%	0.36%	1.10%	-0.04%	-0.34%	0.00%	1.96%	-0.12%	0.93%	2.85%	1.85%	2.11%	-1.96%	-0.97%	-0.03%	0.28%	1.18%	-0.36%

Figura 29. OPUS-Rota χ_1(%) - SCWRL4 χ_1(%) por tipo de aminoácido para diferentes umbrales.



	ALL	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Conjunto 379	89.3%	81.8%	90.1%	88.8%	92.7%	84.6%	78.3%	91.1%	98.6%	95.4%	81.9%	89.0%	96.6%	88.2%	75.8%	94.0%	93.0%	95.6%	97.1%
Conjunto 770	85.6%	79.7%	83.8%	84.6%	89.5%	79.5%	73.2%	87.7%	95.6%	92.3%	76.8%	84.8%	94.3%	86.5%	71.4%	90.0%	92.4%	93.9%	92.9%

Figura 30. Medidas de calidad χ_1(%) por tipo de aminoácido utilizando el método SCWRL4 para dos conjuntos de pruebas diferentes. El conjunto 379 (Krivov *et al.*, 2009) y el conjunto 770 (nuestro conjunto).



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
SCWRL4	0.0526	2.3786	0.953	0.8214	0.3709	1.5052	1.4693	0	1.2066	0.3949	0.509	1.5566	1.1183	1.2193	0.2219	0.5705	0.342	1.326	1.295	0.2757
OPUS-Rota	0.0526	2.3282	0.9495	0.8236	0.3444	1.443	1.4983	0	1.1463	0.3949	0.5024	1.5024	0.9843	1.1422	0.2357	0.5844	0.3417	1.2034	1.241	0.2903

Figura 31. RMSD promedio por tipo de aminoácido.

residuos que no tienen forma alargada, sino anillada, como es el caso de la fenilalanina (PHE), tirosina (TYR), entre otros.

Además de la alanina (ALA) y la glicina (GLY), los aminoácidos que tienen menor RMSD son la prolina (PRO), cisteína (CYS), valina (VAL) y treonina (THR). Estos aminoácios, tienen sólo dos o tres átomos en su cadena lateral, ocacionando que el RMSD sea bajo.

Calidad por caso

En las Figuras 32-35 se muestra la diferencia de las medidas de exactitud absoluta que se obtuvieron para ambos métodos por caso. Se puede observar para las medidas ($\chi_1(\%)$, $\chi_{1,2}(\%)$ y $\chi_{1,2,3}(\%)$), en la mayoría de los casos, la diferencia (OPUS-Rota - SCWRL4) es positiva, es decir, el método OPUS-Rota mejoró las predicciones del SCWRL4.

Para la medida $\chi_1(\%)$, en 434/770 casos, la diferencia de ambos métodos fue positiva, mientras que para las demás medidas $\chi_{1,2}(\%)$, $\chi_{1,2,3}(\%)$ y $\chi_{1,2,3,4}(\%)$ ocurrió para 418/770, 393/770 y 373/770 casos, respectivamente.

En las Figuras 36(a) - 36(d) se muestran las distribuciones de las diferencias de las medidas OPUS-Rota $\chi_1(\%)$ y SCWRL4 $\chi_1(\%)$, hasta OPUS-Rota $\chi_{1,2,3,4}(\%)$ y SCWRL4 $\chi_{1,2,3,4}(\%)$, junto con la distribución normal ajustada a los datos. Basándose en el teorema del límite central, los datos siguen una distribución normal, y en la

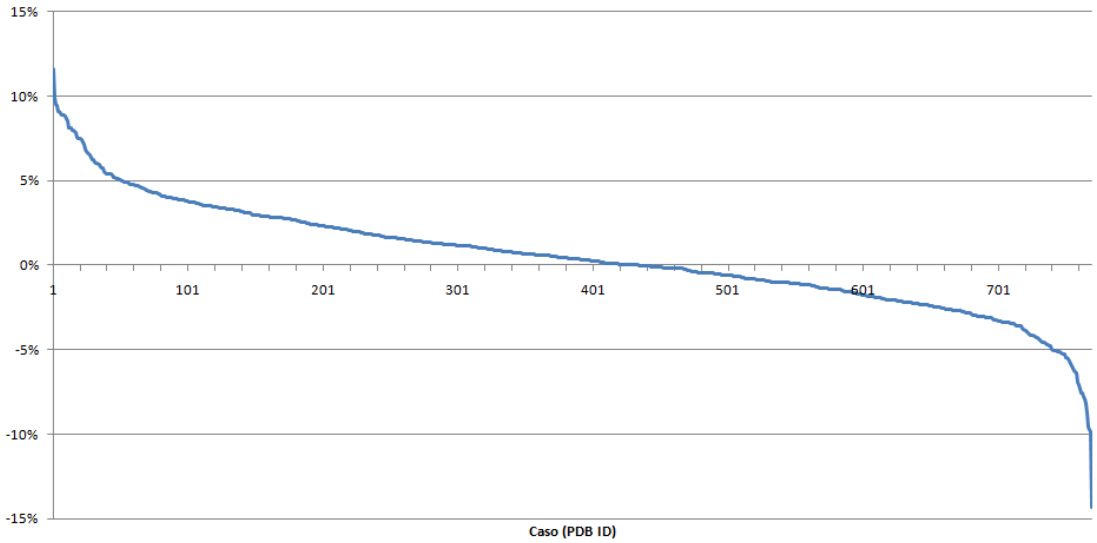


Figura 32. OPUS-Rota χ_1(%) - SCWRL4 χ_1(%)

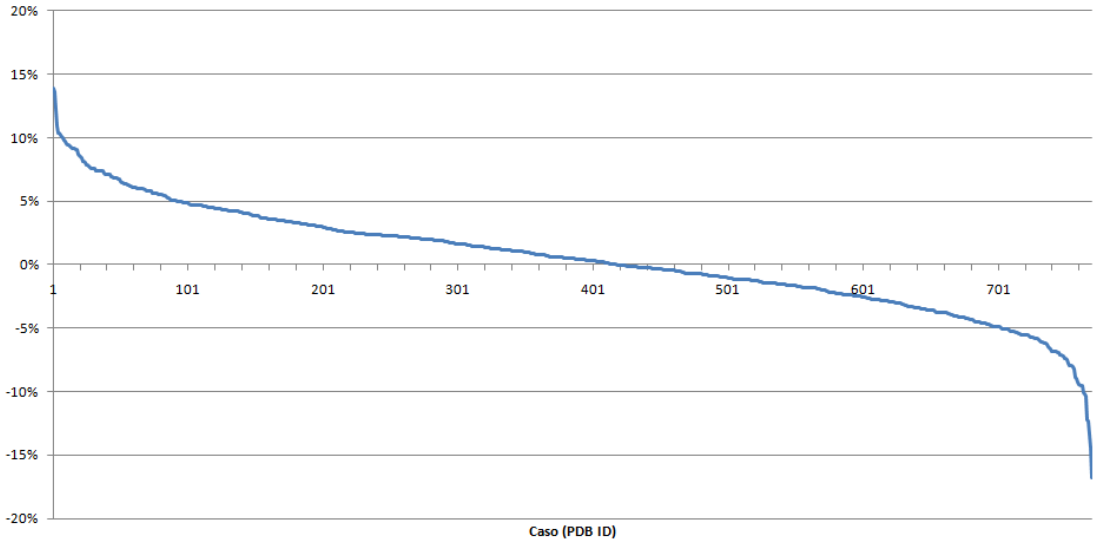


Figura 33. OPUS-Rota $\chi_{1,2}$(%) - SCWRL4 $\chi_{1,2}$(%)

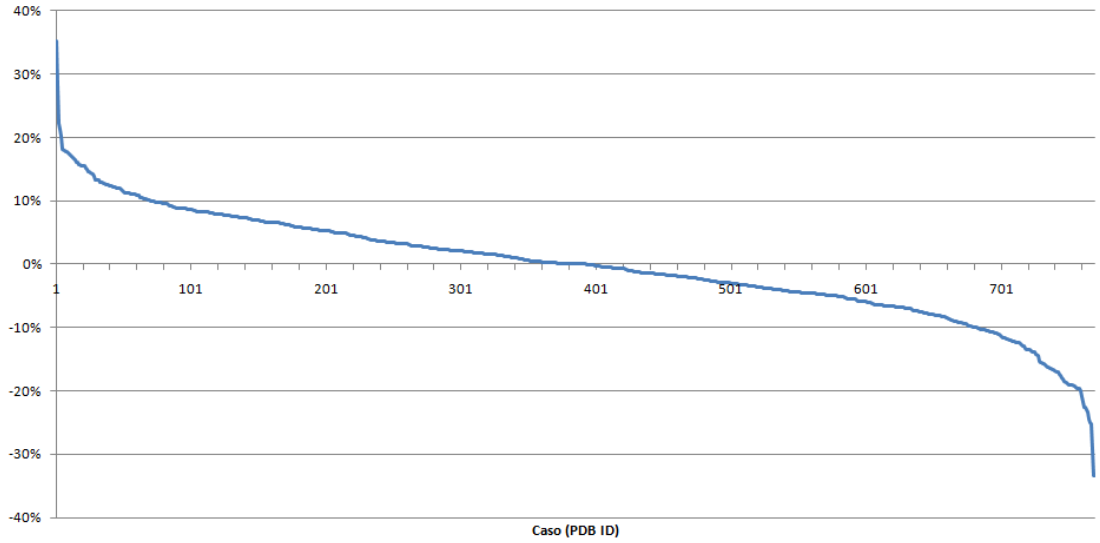


Figura 34. OPUS-Rota $\chi_{1,2,3}$(%) - SCWRL4 $\chi_{1,2,3}$(%)

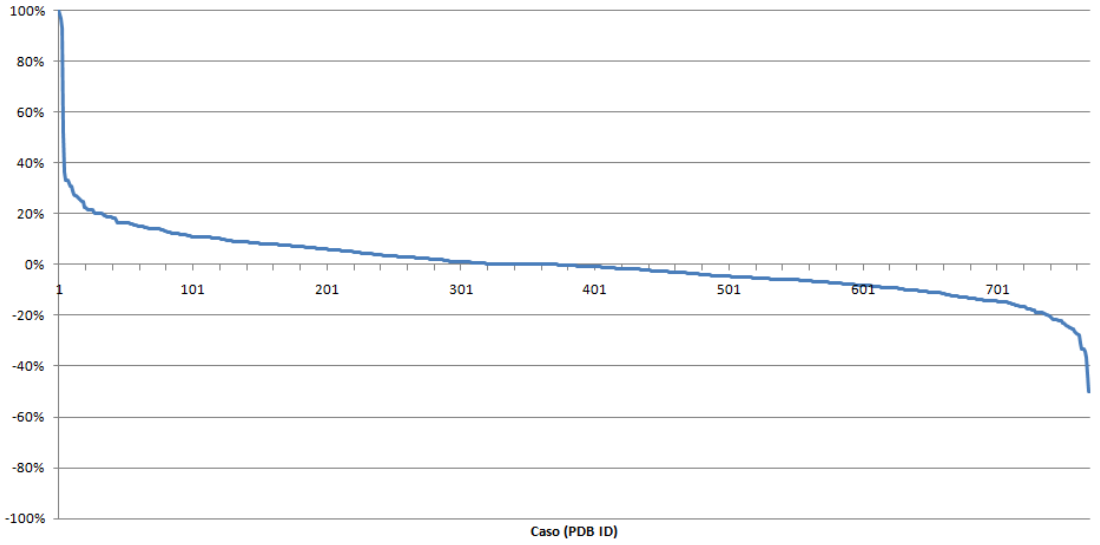


Figura 35. OPUS-Rota $\chi_{1,2,3,4}$(%) - SCWRL4 $\chi_{1,2,3,4}$(%)

Tabla XIII. Intervalos de confianza de la media poblacional de las diferecias de las medidas $\chi_1(\%)$, $\chi_{1,2}(\%)$, $\chi_{1,2,3}(\%)$ y $\chi_{1,2,3,4}(\%)$ de ambos métodos con 99% de confianza.

Medida	Intervalo de Confianza		Método con mayor exactitud
OPUS-Rota $\chi_1(\%)$ - SCWRL4 $\chi_1(\%)$	0.2066	0.7887	OPUS-Rota
OPUS-Rota $\chi_{1,2}(\%)$ - SCWRL4 $\chi_{1,2}(\%)$	-0.3256	0.3367	Indefinido
OPUS-Rota $\chi_{1,2,3}(\%)$ - SCWRL4 $\chi_{1,2,3}(\%)$	-2.3062	-0.7441	SCWRL4
OPUS-Rota $\chi_{1,2,3,4}(\%)$ - SCWRL4 $\chi_{1,2,3,4}(\%)$	-1.2416	1.0578	Indefinido

Tabla XIII se muestran los intervalos de confianza de las medias poblacionales de las diferencias de las medidas de exactitud absoluta de ambos métodos.

Tiempo de ejecución

En la Figura 37 se muestra la distribución del tiempo de ejecución[3] de los métodos SCWRL4 y OPUS-Rota. El tiempo máximo registrado para el método SCWRL4 es de 140.05 segundos, mientras que para el método OPUS-Rota, el tiempo máximo registrado es de 23.06 segundos. Para 58 de las 770 secuencias, el método SCWRL4 tardó más de 23.06 segundos, que fue el tiempo máximo del OPUS-Rota. En la Figura 37 se puede observar que el método SCWRL4 tarda, para la mayoría de sus casos, menos tiempo que el OPUS-Rota; sin embargo, el método SCWRL4 dura en el orden de minutos para varios casos (23/770 tardan más de un minuto en cada caso). El tiempo promedio para el método SCWRL4 y OPUS-Rota es de 9.62 s y 8.66 s, respectivamente.

En el SCWRL4, el tiempo de ejecución depende de la complejidad del caso. Por otro lado, el tiempo de ejecución del OPUS-Rota es un factor que se puede manipular, ya que depende de los parámetros del algoritmo, y no de la entrada.

Máxima exactitud

El objetivo de este experimento es mostrar las medidas de exactitud absoluta máximas que se puede alcanzar al discretizar el espacio de búsqueda mediante el uso de una bilioteca de rotámeros específica.

En la Figura 38 se muestra la máxima exactitud alcanzada utilizando la biblioteca de rotámeros independiente (Dunbrack y Cohen, 1997) para el conjunto de pruebas. Se tomaron las 770 proteínas, y por cada residuo, se buscó en la biblioteca de rotámeros

[3]Los programas se ejecutaron en una computadora con procesador Intel ®Core ™2 Duo @ 2.00 GHz, con 2GB de RAM y sistema operativo Ubuntu de 32-bits.

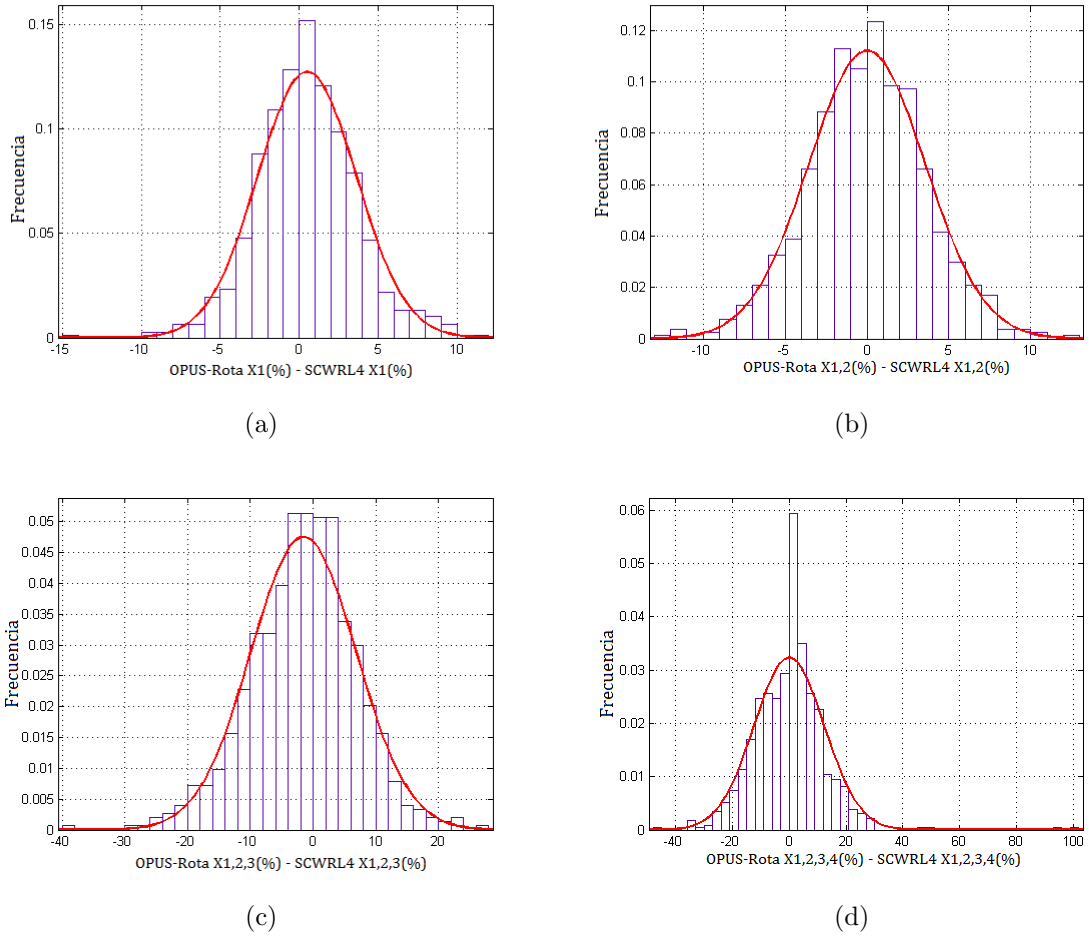


Figura 36. Histograma de las diferencias de las medidas de exactitud absoluta para los métodos OPUS-Rota y SCWRL4.

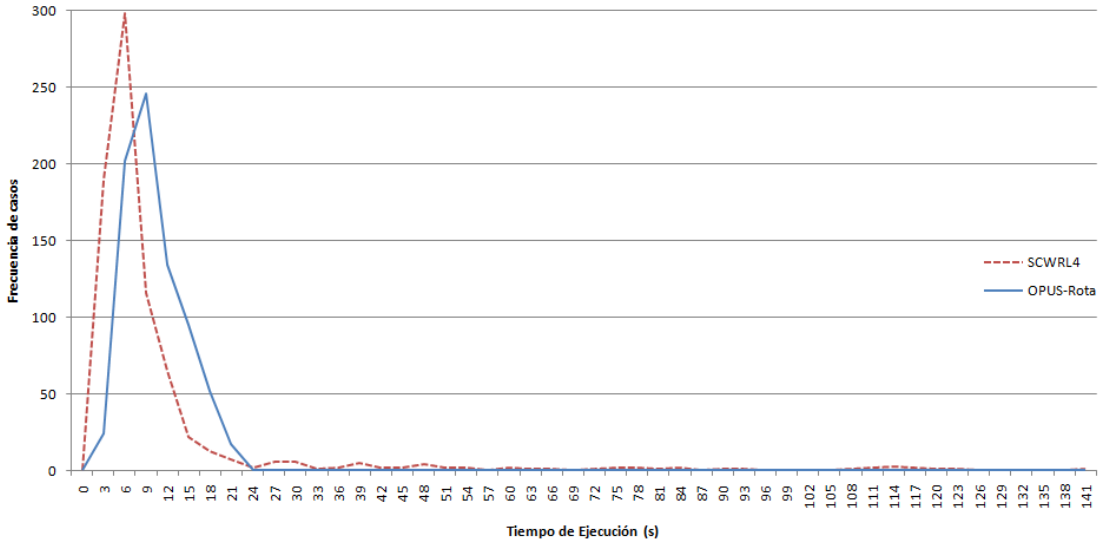
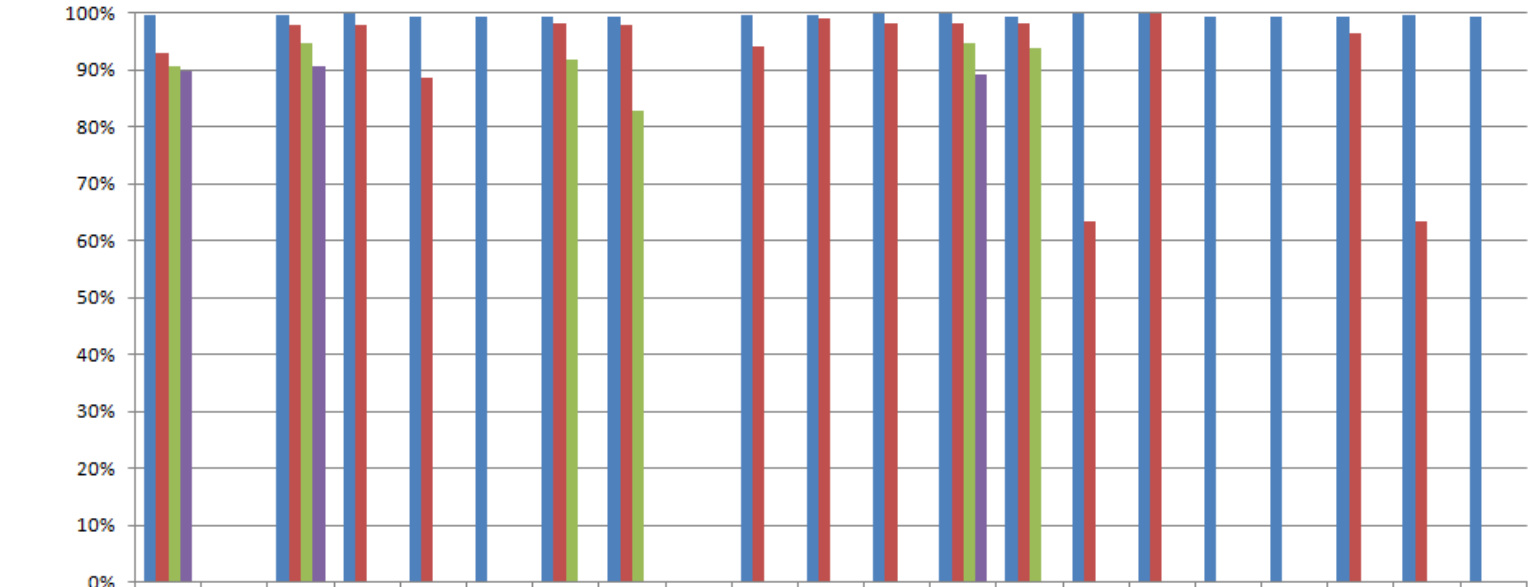


Figura 37. Comparación del tiempo de ejecución de los métodos SCWRL4 y OPUS-Rota.

aquel que estuviera más cerca de los ángulos originales. Al finalizar se obtuvieron las medidas $\chi_1(\%)$, $\chi_{1,2}(\%)$, $\chi_{1,2,3}(\%)$, $\chi_{1,2,3,4}(\%)$.

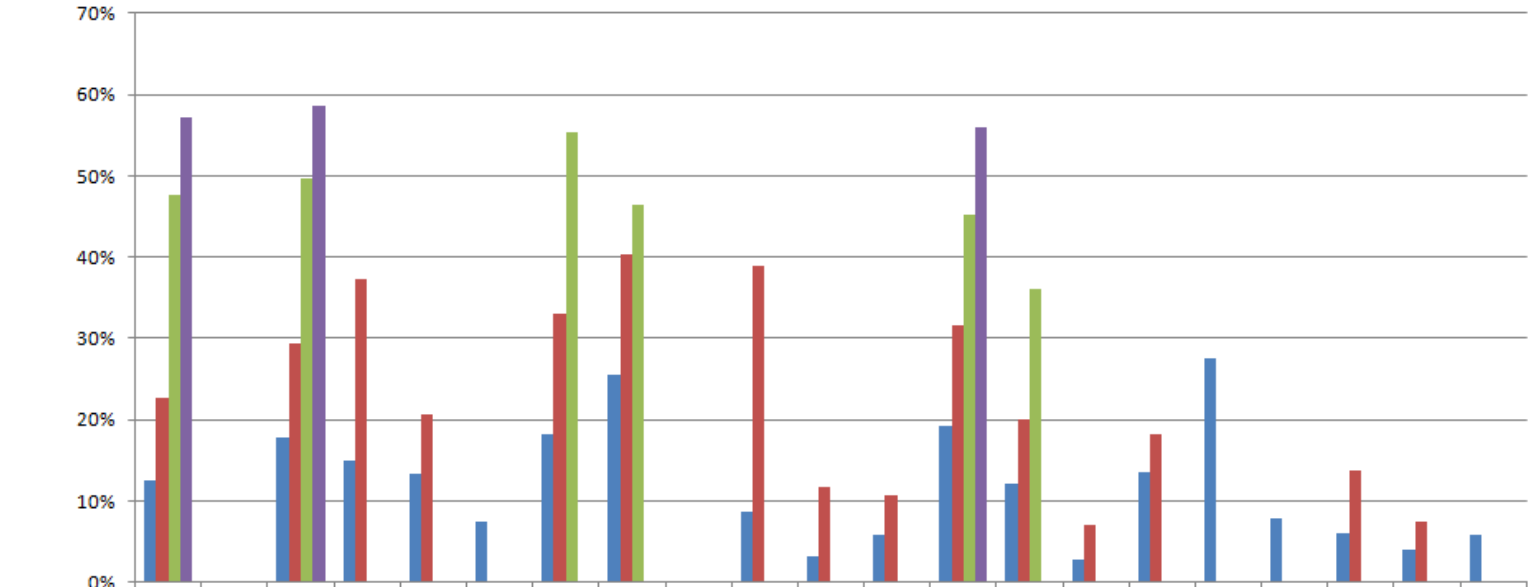
Se puede observar que la medida $\chi_1(\%)$ rebasa el 99% para todos los tipos de aminoácidos, lo que quiere decir que aún si los métodos que aproximan el PSCPP, se limitan a los rotámeros disponibles en la biblioteca de rotámeros independiente, estos métodos pueden alcanzar el 99% de exactitud en la medida $\chi_1(\%)$ con umbral de 40^o. Sin embargo, se ve que los mejores métodos actuales aun están por debajo del 90% de exactitud en la mayoría de los tipos de aminoácidos, por lo que hay un gran margen de mejora en estos métodos.

En la Figura 39 se muestra la diferencia entre la máxima exactitud que se muestra en la Figura 38 y la mejor exactitud realizada por los métodos OPUS-Rota y SCWLR4. En la Tabla XII se puede observar que de manera global, la mejor predicción con umbral de 40^o la hizo el método OPUS-Rota con una medida de exactitud absoluta $\chi_1(\%)$ de 87.10%, mientras que la máxima exactitud es de 99.60%, y la diferencia entre ellas es de 12.50% que es el margen de mejora para los métodos que aproximan el PSCPP.



	ALL	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
χ1(%)	99.6%	0.0%	99.7%	99.8%	99.5%	99.3%	99.4%	99.2%	0.0%	99.6%	99.7%	99.9%	99.8%	99.4%	99.8%	100.0%	99.3%	99.3%	99.5%	99.7%	99.4%
χ1,2(%)	93.0%	0.0%	98.0%	97.9%	88.6%	0.0%	98.1%	97.9%	0.0%	94.1%	99.1%	98.1%	98.1%	98.3%	63.4%	100.0%	0.0%	0.0%	96.4%	63.5%	0.0%
χ1,2,3(%)	90.8%	0.0%	94.7%	0.0%	0.0%	0.0%	91.8%	82.9%	0.0%	0.0%	0.0%	0.0%	94.7%	94.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
χ1,2,3,4(%)	89.8%	0.0%	90.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	89.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figura 38. Medidas de exactitud absoluta máxima global (ALL) y por tipo de aminoácido (códigos de tres letras), alcanzada para el conjunto de pruebas de 770 proteínas utilizando una biblioteca de rotámeros independiente de la columna vertebral.



	ALL	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
χ1(%)	12.5%	0.0%	17.8%	15.0%	13.3%	7.4%	18.2%	25.5%	0.0%	8.6%	3.2%	5.9%	19.1%	12.2%	2.9%	13.5%	27.5%	7.9%	6.1%	4.1%	5.8%
χ1,2(%)	22.7%	0.0%	29.3%	37.3%	20.6%	0.0%	33.0%	40.3%	0.0%	38.8%	11.8%	10.7%	31.6%	20.0%	7.1%	18.2%	0.0%	0.0%	13.7%	7.6%	0.0%
χ1,2,3(%)	47.6%	0.0%	49.6%	0.0%	0.0%	0.0%	55.4%	46.5%	0.0%	0.0%	0.0%	0.0%	45.2%	36.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
χ1,2,3,4(%)	57.2%	0.0%	58.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	56.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figura 39. Diferencia entre la máxima exactitud global (ALL) y por tipo de aminoácido (códigos de tres letras), y la mejor medida de exactitud absoluta alcanzada por los métodos SCWRL4 y OPUS-Rota.

IV.4 Discusión

IV.4.1 Conjunto de pruebas

Los experimentos que se realizaron sobre el conjunto de pruebas mostraron que éste tiene carcterísticas representativas del universo de proteínas conocidas para poderse utilizar como base para comparaciones posteriores entre nuevos métodos que surjan.

Esto es muy importante, ya que los nuevos métodos que aproximen el PSCPP pueden presentar sus resultados utilizando este conjunto como estándar, permitiendo una comparación más sencilla entre previos métodos, es decir, no necesitarían tener los programas de los otros métodos, ni implementarlos, sino sólamente hacer los experimentos con sus propios métodos.

Esta ventaja se aprovecharía más aún si se encontraran disponibles, de manera abierta, las estructuras predichas para todos los casos del conjunto de pruebas, de ambos métodos. Así los grupos que quieran hacer otro tipo de comparaciones, agrupando los residuos de una u otra manera, tengan acceso a las estructuras y de ahí hacer sus comparaciones, como se dijo anteriormente, sin la necesidad de obtener las estructuras nuevamente con los métodos previos.

Sería interesante hacer otro conjunto de casos de prueba, permitiendo proteínas con más de una cadena, inclusive un conjunto para proteínas con más de 400 aminoácidos (proteínas grandes) para evaluar el desempeño de los métodos para estos casos.

IV.4.2 Comparación de métodos

Exactitud

Respecto a la comparación de los métodos, se puede inferir que los métodos no son sensibles a la clasificación SCOP, ni a la clasificación EC; sin embargo, el tipo de aminoácido sí afecta la exactitud.

Esto se viene presentanto en varios métodos, y no es un problema del método, sino que las propiedades físico-químicas de los aminoácidos hacen que estos sean fáciles o difíciles de predecir. Por lo tanto, es importante considerar en el método el tipo de aminoácido que se está manipulando. Es conveniente notar que los esfuerzos de abstraer las propiedades físico-químicas de los aminoácidos para generar mejores predicciones no han sido suficientes.

Máxima exactitud

Se mostró que ambos métodos tienen un margen grande de mejora. Los aspectos que pueden estar afectando a los métodos para no alcanzar la máxima exactitud es la biblioteca de rotámeros y la función objetivo.

La máxima exactitud presentada en los experimentos, se obtuvo con una biblioteca de rotámeros considerada hoy en día como obsoleta. Sin embargo, se ve que la máxima exactitud de esta biblioteca está arriba del 99% en la mayoría de los casos, así que se puede esperar que las nuevas bibliotecas permitan una exactitud del 100%, mientras que los métodos SCWRL4 y OPUS-Rota apenas alcanzan el 85% aproximadamente.

Tiempo de ejecución

Los tiempos de ejecución promedio de los métodos son muy cercanos. En los experimentos se pudo observar que el SCWRL4 puede llegar a tener tiempos de ejecución por caso muy elevados, de hasta 140 segundos. Esto nos dice que el tiempo de ejecución del SCWRL4 depende del caso, mientras que el OPUS-Rota fue más consistente, por lo que el tiempo de ejecución es menos dependiente del caso. Esta independencia se presenta debido a que el OPUS-Rota se basa en la técnica de recocido simulado, por lo que el tiempo de ejecución es una variable del algoritmo, y no del caso de entrada.

Capítulo V

CONCLUSIONES Y PERSPECTIVAS DE INVESTIGACIÓN

En este capítulo se presentan las conclusiones a las que se llegó con la realización de los experimentos aquí presentados, así como propuestas para mejorar la calidad de los algoritmos analizados.

V.1 Sumario

El problema conocido como empaquetamiento de la cadena lateral en proteínas (PSCPP) se reduce a un problema de optimización combinatoria. Para esta reducción se proponen: funciones de energía, que aproximan el plegamiento de las proteínas; bibliotecas de rotámeros, que es un método de discretización de los ángulos de torsión de las cadenas laterales; y métodos de búsqueda para encontrar el mínimo global de la función de energía propuesta.

Bajo esta definición, Akutsu (1997) demostró que el problema pertenece a la clase NP-difícil, para la cual, el uso de heurísticas es una alternativa comúnmente usada.

En la actualidad se tienen las dos mejores heurísticas que aproximan el PSCPP (SCWRL4 y OPUS-Rota) y que difieren en los tres aspectos escenciales: función de energía, biblioteca de rotámeros y método de búsqueda.

El método SCWRL4 es determinístico, mientras que el OPUS-Rota se basa en la técnica de recocido simulado, por lo que es estocástico. Estos métodos no se habían comparado entre sí, por lo que se plantea la pregunta de cuál método es mejor, y bajo qué condiciones.

El método SCWRL3, lo utilizan ampliamente otros métodos de predicción de estructura (Canutescu *et al.*, 2003; Wang *et al.*, 2008). Sin embargo, su popularidad no asegura que su calidad sea superior a la de otros métodos que aproximan el PSCPP.

Ya se mostró experimentalmente (Lu *et al.*, 2008b; Krivov *et al.*, 2009) que los métodos SCWRL4 y OPUS-Rota son mejores que el SCWRL3. Más aún, debido a que

el SCWRL4 es la siguiente versión del SCWRL3 se puede pensar que el SCWRL4 es el mejor candidato para reemplazar al SCWRL3, aunque hay otros métodos como el OPUS-Rota que también se deben considerar.

A continuación se explican las conclusiones a las que se llegó en base a los experimentos realizados en el presente trabajo de investigación.

V.2 Conclusiones

La media poblacional de la diferencia de la medida $\chi_1(\%)$ del método OPUS-Rota y SCWRL4 (OPUS-Rota $\chi_1(\%)$ - SCWRL4 $\chi_1(\%)$) está entre 0.2066 y 0.7887 con 99% de confianza. Este intervalo indica la cercanía que existe entre las soluciones generadas por ambos métodos.

El tiempo de ejecución es otro criterio importante en el que ambos métodos son similares, ya que aunque el método SCWRL4 es un poco más rápido en la mayoría de los casos, éste puede llegar a tardar 7 veces más que el tiempo máximo registrado para el OPUS-Rota.

La clasificación SCOP no es determinante para decidir la calidad de solución en ninguno de los métodos, así como tampoco lo es la clase EC. Sin embargo, el tipo de aminoácido sí es un factor relevante para la calidad de la predicción, es decir, para proteínas con secuencias con alta densidad de prolinas y serinas, se puede recomendar usar el SCWRL4, mientras que para proteínas donde abunden los aminoácidos histidina, lisina, metionina o fenilalanina, es mejor usar el OPUS-Rota, pero para todos los demás residuos no existe un claro ganador.

Las medidas utilizadas, exactitud absoluta y RMSD, muestran aspectos muy diferentes, ya que como se puede observar en las figuras 28, 29 y 31, cuando se tiene el ejemplo de la prolina (PRO) en la que el método SCWRL4 parece generar mejores predicciones, usando la medida de exactitud absoluta, esto no se ve reflejado de la misma manera en el RMSD. Esto motiva a encontrar otra manera de medir la calidad de las predicciones para el PSCPP.

V.3 Perspectivas de investigación

V.3.1 Implementación de un algoritmo y pruebas adicionales

El OPUS-Rota tiene la ventaja de basarse en la técnica de recocido simulado, por lo que es fácil modificar: la función objetivo, la biblioteca de rotámeros, y otros parámetros como la función de vecindario. Esta flexibilidad proporciona una oportunidad de modificar el método para tratar de mejorarlo.

Entonces, se propone implementar un método basado en recocido simulado, parecido al OPUS-Rota, para hacer varias combinaciones con diferentes funciones objetivo, bibliotecas de rotámeros y funciones de vecindario, para analizar la combinación que genere mejores resultados.

La biblioteca de rotámeros utilizada para el método SCWRL4 es la más actual; aunque aún no está disponible. La idea es utilizar esta biblioteca de rotámeros cuando se encuentre disponible para analizar si hubo mejoras en el método OPUS-Rota. Además, el método SCWRL4 utiliza el concepto de subrotámeros, que consiste en utilizar como rotámeros, aquellos que esten a una distancia de $\pm\delta$ (desviación estándar) de cada rotámero de la biblioteca utilizada. Si bien este concepto representa el tener que trabajar con un espacio de búsqueda mayor, el método de recocido simulado puede ayudar a minimizar el tiempo de búsqueda.

Para la función de vecindario, se propone tener dos funciones de vecindario, que generen conjuntos de vecindario mutuamente excluyentes, de tal manera que se ejecute el recocido simulado utilizando una función, y al terminar, ejecutarlo de nuevo pero ahora con la otra función de vecindario, y así sucesivamente. Para esto se recomienda analizar la técnica conocida como vecindarios de tamaño variable (VNS) (Mladenovic y Hansen, 1997).

Para la función de energía se propone estudiar y analizar las funciones utilizadas por el SCWRL4 y el OPUS-Rota, y tratar de encontrar una función que contemple lo mejor de ambas.

Se recomienda hacer un análisis empleando otros parámetros de las proteínas, como la superficie expuesta al solvente. Este parámetro está relacionado al tipo de aminoácido; sin embargo, puede aportar más información en el caso de aminoácidos que están tanto en el centro como en la superficie de las proteínas.

También se recomienda hacer experimentos con proteínas con más de una cadena,

para analizar si esto afecta o no en la calidad de las soluciones.

V.3.2 Medidas de calidad

Además de implementar un nuevo método, se propone definir otra medida de calidad que sea continua como el RMSD, pero que se base en ángulos como las medidas de exactitud absoluta y condicional.

El problema de determinar la similitud entre un par de estructuras se viene tratando de resolver de diferentes maneras, utilizando diferentes medidas. El RMSD es una métrica ampliamente utilizada para determinar la similitud entre estructuras; sin embargo, cuando se aplica al PSCPP tiene algunas limitaciones.

El RMSD se basa en las coordenadas de los átomos, y en los métodos que aproximan el PSCPP, el espacio de búsqueda son los ángulos de torsión de las cadenas laterales y no las coordenadas.

Se puede tener una conformación, cuyos ángulos sean idénticos a los de la conformación nativa; pero puede tener un RMSD mayor que cero. Esto debido a que los métodos que resuelven el PSCPP utilizan distancias promedio de los enlaces covalentes, para pasar del espacio de ángulos de torsión a coordenadas.

Por el contrario, las medidas basadas en los ángulos de torsión, como lo son las medidas de exactitud absoluta y condicional son medidas discretas, es decir, se propone un umbral, y se cuenta como ángulo correcto aquel cuyo error esté dentro del umbral especificado.

Entonces, se puede tener un par de conformaciones predichas, ambas con la misma calidad bajo estas medidas de exactitud, pero donde una, puede tener todos los ángulos exactamente igual que la conformación nativa, es decir, que el error de los ángulos sea cero, mientras que la otra conformación, puede tener el error de los ángulos justo por debajo del umbral. Se puede pensar que la mejor conformación es la primera, aquella cuyos ángulos están más cerca de la conformacón nativa, pero la medida de exactitud no distingue entre estas dos conformaciones y ambas tendrían la misma calidad, debido a que la medida es discreta.

Por consiguiente, se ve la necesidad de crear una medida que sea continua como el RMSD, pero que esté midiendo el error de los ángulos de torsión de la cadena lateral.

Referencias

(2010). RCSB protein data bank. http://www.pdb.org/pdb/home/home.do.

(2010a). UniProtKB/Swiss-Prot release 2010_09 statistics. http://ca.expasy.org/sprot/relnotes/relstat.html.

(2010b). UniProtKB/TrEMBL release statistics | UniProt | the universal protein resource | EBI. http://www.ebi.ac.uk/uniprot/TrEMBLstats/.

Akutsu, T. (1997). NP-Hardness results for protein Side-Chain packing. *Genome Inform Ser*, (8): 180–186.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., y Tasumi, M. (1978). The protein data bank: A computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, **185**(2): 584–591.

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies*, **55**(4): 533–552.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., y Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.*, **31**(1): 365–370.

Bower, M. J., Cohen, F. E., y Dunbrack, R. L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *Journal of Molecular Biology*, **267**(5): 1268–1282. PMID: 9150411.

Brenner, S. E. (2009). Scop parseable files. http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html.

Bryant, S. H. y Altschul, S. F. (1995). Statistics of sequence-structure threading. *Current Opinion in Structural Biology*, **5**(2): 236–244.

Canutescu, A. A., Shelenkov, A. A., y Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science: A Publication of the Protein Society*, **12**(9): 2001–2014. PMID: 12930999.

Chazelle, B., Kingsford, C., y Singh, M. (2004). A semidefinite programming approach to side chain positioning with new rounding strategies. *INFORMS J. on Computing*, **16**(4): 380–392.

Dunbrack, J. R. L. (2002). Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, **12**(4): 431–440. PMID: 12163064.

Dunbrack, R. L. y Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science: A Publication of the Protein Society*, **6**(8): 1661–1681. PMID: 9260279.

Dunbrack, R. L. y Karplus, M. (1993). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *Journal of molecular biology*, **230**(2): 574, 543.

Engh, R. A. y Huber, R. (1991). Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallographica Section A Foundations of Crystallography*, **47**(4): 392–400.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., y Baker, D. (2003). Protein-Protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, **331**(1): 281–299.

Grigoryan, G., Ochoa, A., y Keating, A. E. (2007). Computing van der waals energies in the context of the rotamer approximation. *Proteins: Structure, Function, and Bioinformatics*, **68**(4): 863–878.

Gu, J. y Bourne, P. E. (2009). *Structural Bioinformatics*. Wiley-Blackwell. ISBN 9780470181058.

Hsin, J., Yang, C., Huang, K., y Yang, C. (2007). An ant colony optimization approach for the protein side chain packing problem. En *Proceedings of the 6th conference on Microelectronics, nanoelectronics, optoelectronics*, páginas 44–49, Istanbul, Turkey. World Scientific and Engineering Academy and Society (WSEAS). ISBN 978-960-8457-74-4.

Jain, T., Cerutti, D. S., y McCammon, J. A. (2006). Configurational-bias sampling technique for predicting side-chain conformations in proteins. *Protein Science: A Publication of the Protein Society*, **15**(9): 2029–2039. PMID: 16943441.

Janin, J., Wodak, S., Levitt, M., y Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *Journal of Molecular Biology*, **125**(3): 357–386.

Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., ærfeldt, H. S., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A., y Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology*, **319**(5): 1257–1265.

Kingsford, C. L., Chazelle, B., y Singh, M. (2005). Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, **21**(7): 1028–1039.

Kirkpatrick, S., Gelatt, C. D., y Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**(4598): 671–680. ArticleType: research-article / Full publication date: May 13, 1983 / Copyright Â© 1983 American Association for the Advancement of Science.

Krivov, G. G., Shapovalov, M. V., y Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. PMID: 19603484.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armananzas, R., Santafe, G., Perez, A., y Robles, V. (2006). Machine learning in bioinformatics. *Brief Bioinform*, **7**(1): 86–112.

Leja, D. (2010). Enzyme. http://www.accessexcellence.org/RC/VL/GG/enzyme.php.

Liang, S. y Grishin, N. V. (2002). Side-chain modeling with an optimized scoring function. *Protein Science: A Publication of the Protein Society*, **11**(2): 322–331. PMID: 11790842.

Lu, M., Dousis, A. D., y Ma, J. (2008a). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of Molecular Biology*, **376**(1): 288–301.

Lu, M., Dousis, A. D., y Ma, J. (2008b). OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Science*, **17**(9): 1576–1585.

Mladenovic, N. y Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, **24**(11): 1097–1100.

Moss, G. (2010). Enzyme nomenclature. http://www.chem.qmul.ac.uk/iubmb/enzyme/.

Murzin, A. G., Brenner, S. E., Hubbard, T., y Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**(4): 536–540.

Osguthorpe, D. J. (2000). Ab initio protein folding. *Current Opinion in Structural Biology*, **10**(2): 146–152.

Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., y Strauss, C. E. M. (2005). Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of Computational Chemistry*, **26**(10): 1063–1068. PMID: 15898109.

Peterson, R. W., Dutton, P. L., y Wand, A. J. (2004). Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Science: A Publication of the Protein Society*, **13**(3): 735–751. PMID: 14978310.

Platt, O. S., Brambilla, D. J., Rosse, W. F., Milner, P. F., Castro, O., Steinberg, M. H., y Klug, P. P. (1994). Mortality in sickle cell disease. life expectancy and risk factors for early death. *The New England Journal of Medicine*, **330**(23): 1639–1644. PMID: 7993409.

Plaxco, K. W., Simons, K. T., Ruczinski, I., y Baker, D. (2000). Topology, stability, sequence, and length: Defining the determinants of Two-State protein folding kinetics. *Biochemistry*, **39**(37): 11177–11183.

Sánchez, R. y Sali, A. (1997). Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, **7**(2): 206–214.

Shapovalov, M. V. y Dunbrack, R. L. (2007). Statistical and conformational analysis of the electron density of protein side chains. *Proteins: Structure, Function, and Bioinformatics*, **66**(2): 279–303.

Shetty, R. P., Bakker, P. I. W. D., DePristo, M. A., y Blundell, T. L. (2003). Advantages of fine-grained side chain conformer libraries. *Protein Engineering*, **16**(12): 963–969. PMID: 14983076.

Simons, K. T., Kooperberg, C., Huang, E., y Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, **268**(1): 209–225. PMID: 9149153.

Steinbach, P. J. (2005a). The empirical potential energy function. http://cmm.info.nih.gov/intro_simulation/node15.html.

Steinbach, P. J. (2005b). Molecular dynamics (MD) simulation. http://cmm.cit.nih.gov/intro_simulation/node23.html.

Wang, C., Schueler-Furman, O., y Baker, D. (2005). Improved side-chain modeling for protein-protein docking. *Protein Science: A Publication of the Protein Society*, **14**(5): 1328–1339. PMID: 15802647.

Wang, G. y Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics (Oxford, England)*, **19**(12): 1589–1591. PMID: 12912846.

Wang, G. y Dunbrack, R. L. (2010). S2C:A database correlating sequence and atomic coordinate numbering in the protein data bank. http://dunbrack.fccc.edu/Guoli/s2c/.

Wang, Q., Canutescu, A. A., y Dunbrack, R. L. (2008). SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat. Protocols*, **3**(12): 1832–1847.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., y Suzek, B. (2006). The universal protein resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.*, **34**(suppl_1): D187–191.

Xiang, Z. y Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology*, **311**(2): 421–430. PMID: 11478870.

Xu, J. (2005). Rapid protein Side-Chain packing via tree decomposition. En *Research in Computational Molecular Biology*, páginas 423–439.

Xu, J. y Berger, B. (2006). Fast and accurate algorithms for protein side-chain packing. *J. ACM*, **53**(4): 533–557.

Yanover, C., Schueler-Furman, O., y Weiss, Y. (2008). Minimizing and learning energy functions for side-chain prediction. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **15**(7): 899–911. PMID: 18707538.

Zhang, J., Gao, X., Xu, J., y Li, M. (2008). Rapid and accurate protein side chain prediction using local backbone information only.

Zhang, Y. y Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**(4): 702–710. PMID: 15476259.

Apéndice A

IMPLEMENTACIÓN DE ALGORITMO

A.1 Algoritmo

En esta sección se describen los detalles de un algoritmo sencillo basado en la técnica de recocido simulado para el problema conocido como empaquetamiento de la cadena lateral en proteínas (PSCPP).

A.1.1 Definiciones

Se describen primero las variables que se van a utilizar en las secciones posteriores.

Aminoácidos

Sea L el conjunto que representa a los 20 aminoácidos estándar. Entonces

$$\begin{aligned} \mathbf{L} &= \{\text{ALA, CYS, ASP, GLU, PHE, GLY, HIS, ILE, LYS, LEU, MET, ASN,} \\ &\quad \text{PRO, GLN, ARG, SER, THR, VAL, TRP, TYR}\} \end{aligned}$$

Biblioteca de rotámeros

Sea **RL** el conjunto que representa a la biblioteca de rotámeros independiente. Entonces

$$\begin{aligned} \mathbf{RL} &= \{R_{ALA}, R_{CYS}, R_{ASP}, R_{GLU}, R_{PHE}, R_{GLY}, R_{HIS}, R_{ILE}, R_{LYS}, R_{LEU}, R_{MET} \\ &\quad R_{ASN}, R_{PRO}, R_{GLN}, R_{ARG}, R_{SER}, R_{THR}, R_{VAL}, R_{TRP}, R_{TYR}\} \\ R_x &= (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_{|R_x|}) \end{aligned}$$

donde R_x, para un $x \in \mathbf{L}$, es la biblioteca de rotámeros para el aminoácido tipo x y $\mathbf{r}_i \in R_x$ es el i-ésimo rotámero para el aminoácido tipo x.

A.1.2 Entradas

Se sabe que el PSCPP tiene como entrada la estructura primaria, *i.e.*, la secuencia de aminoácidos de la proteína, y las coordenadas tridimensionales de la columna vertebral. En esta sección se definen estas variables de entrada.

Secuencia de aminoácidos

Sea A la variable de entrada que representa la secuencia de aminoácidos, ℓ la longitud de esta secuencia, entonces

$$A = (a_1, a_2, a_3, \ldots, a_\ell)$$

donde $a_i \in \mathbf{L}$ es el tipo de aminoácido del i-ésimo residuo de la cadena.

Coordenadas de la columna vertebral

Sea N, CA, C y O las variables de entrada que representan la secuencia de coordenadas de los átomos (nitrógeno, carbono-α, carbono y oxígeno, respectivamente) de la columna vertebral de la proteína, entonces

$$\begin{aligned} N &= (n_1, n_2, n_3, \ldots, n_\ell) \\ CA &= (ca_1, ca_2, ca_3, \ldots, ca_\ell) \\ C &= (c_1, c_2, c_3, \ldots, c_\ell) \\ O &= (o_1, o_2, o_3, \ldots, o_\ell) \\ n_i, ca_i, c_i, o_i &\in \Re^3 \; \forall \; 1 \leq i \leq \ell \end{aligned}$$

A.1.3 Representación

A continuación se describe la forma en la que se representa una solución en el algoritmo. Para este problema, una solución S es una secuencia de enteros que indican los índices de los rotámeros que se seleccionaron, uno por cada residuo. Es decir,

$$S = (s_1, s_2, s_3, \ldots, s_\ell)$$

donde $1 \leq s_i \leq |R_{a_i}|$.

A.1.4 Función de vecindario

Uno de los aspectos más importantes en un algoritmo de búsqueda es la función de vecindario ($N(\cdot)$) que se utiliza para generar una solución a partir de otra. La función de vecindario se define como sigue.

Sea S una solución, $S' \in N(S) \Leftrightarrow \exists! j \ : \ \forall \ i \neq j, \ s'_i = s_i, \ s'_j = (s_j + 1) \bmod (|R_{a_j}|)$.

Dada esta función de vecindario, una solución S tiene O(ℓ) vecinos. Al final del algoritmo, cuando la probabilidad de aceptar una solución con mayor energía sea casi nula, el algoritmo realizará una búsqueda exahustiva en los vecinos de la solución actual, por lo que una función de vecindario de tamaño lineal favorece en este tipo de búsquedas.

A.1.5 Función de energía

Como se menciona en la Sección III.1.2, existen muchas y diversas funciones que se puede utilizar para representar la energía de una proteína. Por facilidad se propone una función cuya evaluación dependa de información con la que ya se cuenta. A continuación se presenta la función $E(S)$ que representa la energía de una solución S (Bower *et al.*, 1997).

$$E(S) = \sum_{i=1}^{\ell} \sum_{j=i}^{\ell} \sum_{a \in BB(i)} \sum_{b \in SC(j)} E(a,b) + \sum_{i=1}^{\ell-1} \sum_{j=i+1}^{\ell} \sum_{a \in SC(i)} \sum_{b \in AT(j)} E(a,b)$$

$$E(a,b) = \begin{cases} 0 & d(a,b) \geq r(a) + r(b) \\ 10 & d(a,b) \leq 0.8254(r(a) + r(b)) \\ 57.273(1 - \frac{d(a,b)}{r(a)+r(b)}) & otherwise \end{cases}$$

donde $BB(i)$ es el conjunto de átomos pesados de la columna vertebral del residuo i ($BB(i) = \{n_i, ca_i, c_i, o_i\}$), $SC(i)$ es el conjunto de átomos pesados de la cadena lateral del residuo i (debido que los aminoácidos varian en la cadena lateral, el tamaño del conjunto $SC(i)$ se determina por cada a_i, es decir, por el tipo de aminoácido). $AT(i) = SC(i) \cup BB(i)$ es el conjunto de todos los átomos pesados del residuo i.

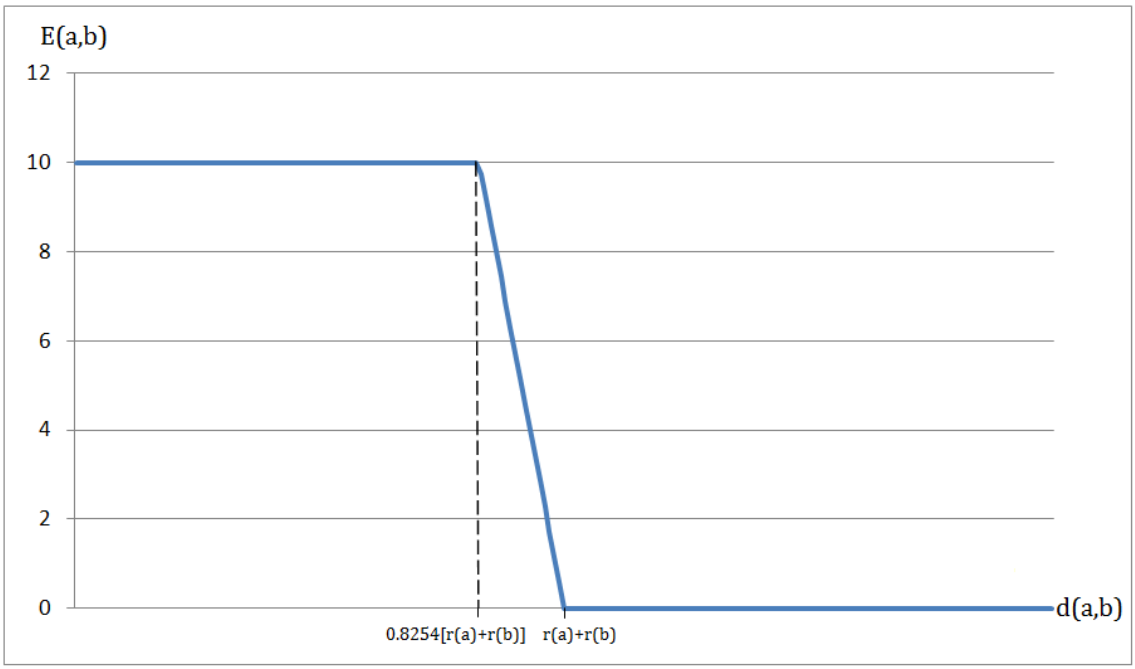


Figura 40. Energía $E(a,b)$ entre los átomos a y b con radios $r(a)$ y $r(b)$ respectivamente que están a una distancia $d(a,b)$.

Es decir, para cada par de átomos (a,b) tal que a, b o ambos sean átomos de la cadena lateral de algún residuo, se calcula el término $E(a,b)$ y la suma es el valor de la energía para la solución dada.

Esta energía se requiere minimizar. La función $E(S)$ no tiene términos negativos, por lo que para minimizarla se requiere que la mayoría de los términos $E(a,b)$ sean lo más cercano a cero posible. En la Figura 40 se muestra la función $E(a,b)$ en función de la distancia $d(a,b)$. Se puede observar que mientras la distancia entre un par de átomos sea mayor que la suma de sus radios, $E(a,b)$ será cero. Por lo que al minimizar $E(S)$ se está buscando una estructura cuyos átomos esten lo más alejados unos de otros, teniendo como restricción las distancias de los enlaces covalentes que existen entre algunos pares de ellos. Viéndolo desde otro punto de vista, se puede decir que la función $E(a,b)$ penaliza los choques interatómicos, es decir, que dos átomos estén ocupando el mismo espacio. Se puede observar entonces, que el algoritmo tratará de encontrar la solución que genere la menor cantidad de choques interatómicos posibles; sin embargo, no asegura que se entregue como resultado una solución sin choques interatómicos.

De ángulos de torsión a coordenadas

En la Sección A.1.3 se describe la representación de una solución, la cual está dada como un conjunto de rotámeros, es decir, un conjunto de ángulos de torsión tomados de la biblioteca de rotámeros. En la Sección A.1.5 se establece la función de energía, en la cual se necesita calcular las distancias entre pares de átomos. Uno se puede dar cuenta que existe una discrepancia en esto, ya que la representación está dada en ángulos, y se necesitan las coordenadas de los átomos de las cadenas laterales para calcular la energía de una solución.

El problema entonces consiste en convertir la información de los ángulos de torsión en coordendas. Al analizar este problema, uno se da cuenta que la información de los ángulos de torsión no es suficiente para generar las coordenadas, sino que se necesita información adicional. Esta información se refiere a restricciones las cuales están dadas por los enlaces covalentes que existen en las cadenas laterales y de acuerdo con los átomos que formen el enlace es la distancia y el ángulo que forma el enlace.

A continuación se define el problema de manera general como un problema de geometría.

Dados los puntos $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \Re^3$, los ángulos θ y φ, y la constante $R \in \Re$, determinar el punto $\mathbf{D} \in \Re^3$ tal que cumpla con las siguientes condiciones.

Sean

$\mathbf{AB} \equiv \mathbf{B} - \mathbf{A}$

$\mathbf{BC} \equiv \mathbf{C} - \mathbf{B}$

$\mathbf{CD} \equiv \mathbf{D} - \mathbf{C}$

$\mathbf{n}_1 \equiv \mathbf{AB} \times \mathbf{BC}$

$\mathbf{n}_2 \equiv \mathbf{BC} \times \mathbf{CD}$

Las condiciones que se deben cumplir son:

$$R = |\mathbf{CD}| \tag{25}$$

$$\cos\theta = \frac{\mathbf{BC} \cdot \mathbf{CD}}{|\mathbf{BC}||\mathbf{CD}|} \tag{26}$$

$$\cos\varphi = \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{|\mathbf{n}_1||\mathbf{n}_2|} \tag{27}$$

En la Figura 41 se muestra un diagrama que ejemplifica el problema.

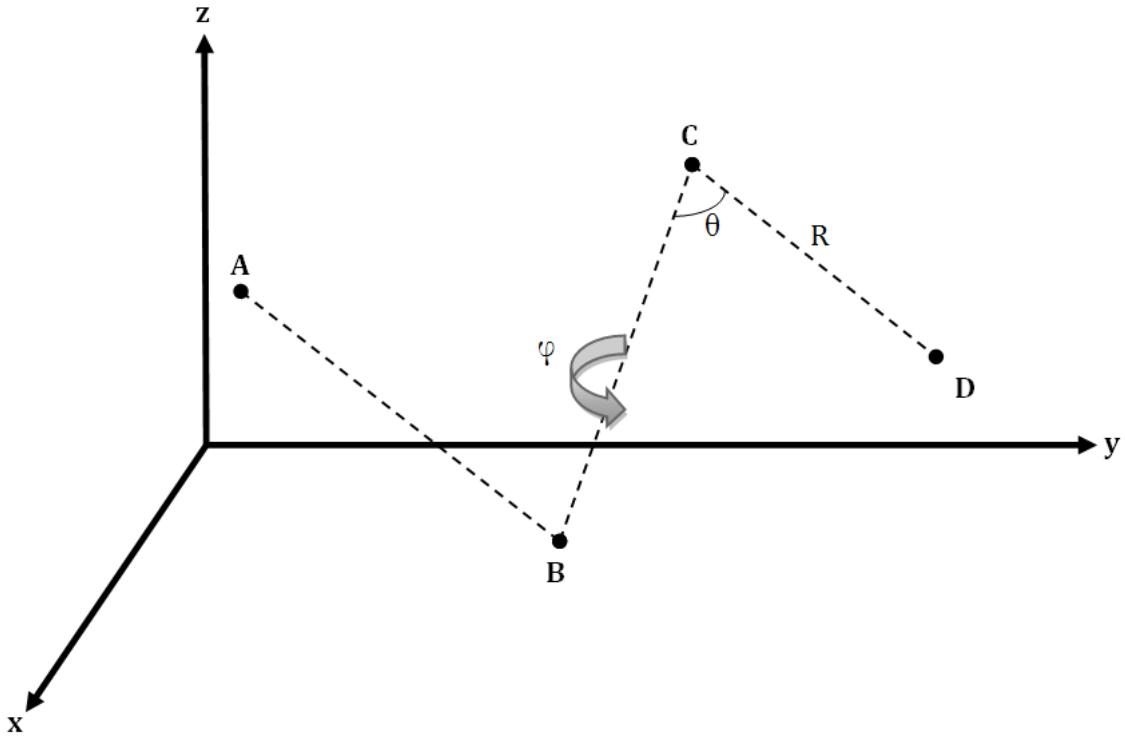


Figura 41. Conversión de ángulo de torisón a coordenadas. Se conocen los vectores **A**, **B**, **C**, el valor de R, los ángulos φ y $theta$ y se quiere determinar el valor **D**.

Es decir, el átomo con coordenadas en **C** está unido al átomo cuyas coordenadas (**D**) se están buscando, pero se conoce la distancia (R) que debe tener el enlace covalente que los une (Ecuación (25)), así como también el ángulo (θ) que deben forman los átomos con coordenadas en **B**, **C** y **D** (Ecuación (26)) y el ángulo de torsión (φ) que forman los cuatro átomos (Ecuación (27)).

En Parsons *et al.* (2005) se describen varios métodos para resolver este problema, de los cuales el método denominado *Natural Extension Reference Frame* (*NeRF*) es el más sencillo de implementar. A continuación se muestra el procedimiento para obtener las coordenadas del punto **D** utilizando el método *NeRF*.

Sea $\mathbf{bc} \equiv \frac{\mathbf{BC}}{|\mathbf{BC}|}$ y $\mathbf{n} \equiv \frac{\mathbf{AB}\times\mathbf{bc}}{|\mathbf{AB}\times\mathbf{bc}|}$. Se genera la matriz **M** de la siguiente manera.

$$\mathbf{M} \equiv [\mathbf{bc}, \mathbf{n} \times \mathbf{bc}, \mathbf{n}] \tag{28}$$

Después se hace $\mathbf{D}_2 = (R\cos\theta, R\cos\varphi\sin\theta, R\sin\varphi\sin\theta)$ con lo que finalmente se obtiene

$$\mathbf{D} = \mathbf{M}\mathbf{D}_2 + \mathbf{C} \tag{29}$$

La idea de este método es primero tomar al punto C como centro de un marco de referencia especial y colocar a $\mathbf{D}_2$ usando las restricciones de R, θ y φ. Posteriormente utilizar a $\mathbf{M}$ para convertir de este marco de referencia al original utilizando ahora la información de $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$.

En Parsons *et al.* (2005) utilizan estos métodos para generar las coordenadas de los átomos de la columna vertebral en algoritmos de predicción de estructura como Rossetta (Simons *et al.*, 1997), pero mencionan que estos métodos se pueden utilizar por igual en las cadenas laterales.

A.1.6 Parámetros

Como se muestra en la Sección A.1.5, para obtener las coordenadas de los átomos de las cadenas laterales a partir de los ángulos de torsión de las bibliotecas de rotámeros, es necesario además, conocer las restricciones dadas por los enlaces covalentes. Estas restricciones se aplican la distancia (R) entre un par de átomos unidos por un enlace covalente y el ángulo θ formado por dos enlaces covalentes contiguos. En la Figura 41 los enlaces covalentes se muestran con las líneas discontinuas. Estos parámetros dependen de los tipos de átomos que forman el enlace, además de otras propiedades químicas.

En Engh y Huber (1991) se muestran tablas que contienen datos de distancias y ángulos de enlaces covalentes promedio que se obtuvieron estadísticamente utilizando estructuras de proteínas conocidas. Estos datos se utilizan ampliamente para convertir los ángulos de torsión a coordenadas.

Para utilizar estos datos fue necesario hacer una correspondencia entre los átomos pesados de las cadenas laterales y una lista de tipo de átomo que se encuentra en Engh y Huber (1991). En la Tabla XIV se muestran los códigos PDB de los átomos pesados de los 20 aminoácidos estándar, mientras que en la Tabla XV se encuentra el tipo de átomo correspondiente de acuerdo a las descripciones de Engh y Huber (1991).

Tabla XIV. Lista de códigos PDB de los átomos pesados por aminoácido.

Aminoácido	Columna Vertebral				Cadena Lateral											
					Átomos Principales					Átomos Redundantes						
A	N	CA	O	C	CB											
C	N	CA	O	C	CB	SG										
D	N	CA	O	C	CB	CG	OD1			OD2						
E	N	CA	O	C	CB	CG	CD	OE1		OE2						
F	N	CA	O	C	CB	CG	CD1			CD2	CE1	CE2	CZ			
G	N	CA	O	C												
H	N	CA	O	C	CB	CG	ND1			CD2	CE1	NE2				
I	N	CA	O	C	CB	CG1	CD1			CG2						
K	N	CA	O	C	CB	CG	CD	CE	NZ							
L	N	CA	O	C	CB	CG	CD1			CD2						
M	N	CA	O	C	CB	CG	SD	CE								
N	N	CA	O	C	CB	CG	OD1			ND2						
P	N	CA	O	C	CB	CG	CD									
Q	N	CA	O	C	CB	CG	CD	OE1		NE2						
R	N	CA	O	C	CB	CG	CD	NE	CZ	NH1	NH2					
S	N	CA	O	C	CB	OG										
T	N	CA	O	C	CB	OG1				CG2						
V	N	CA	O	C	CB	CG1				CG2						
W	N	CA	O	C	CB	CG	CD1			CD2	NE1	CE2	CE3	CZ2	CZ3	CH2
Y	N	CA	O	C	CB	CG	CD1			CD2	CE1	CE2	CZ	OH		

Tabla XV. Códigos de los tipos de átomos de Engh y Huber (1991) en correspondencia con la lista de átomos de la Tabla XIV.

Aminoácido	Columna Vertebral				Cadena Lateral											
					Átomos Principales					Átomos Redundantes						
A	NH1	CH1E	O	C	CH3E											
C	NH1	CH1E	O	C	CH2E	SH1E										
D	NH1	CH1E	O	C	CH2E	C	OC			OC						
E	NH1	CH1E	O	C	CH2E	CH2E	C	OC		OC						
F	NH1	CH1E	O	C	CH2E	CF	CR1E			CR1E	CR1E	CR1E	CR1E			
G	NH1	CH1E	O	CH2G												
H	NH1	CH1E	O	C	CH2E	C5	NR			CR1H	CRHH	NH1				
I	NH1	CH1E	O	C	CH1E	CH2E	CH3E			CH3E						
K	NH1	CH1E	O	C	CH2E	CH2E	CH2E	CH2E	NH3							
L	NH1	CH1E	O	C	CH2E	CH1E	CH3E			CH3E						
M	NH1	CH1E	O	C	CH2E	CH2E	SM	CH3E								
N	NH1	CH1E	O	C	CH2E	C	O			NH2						
P	N	CH1E	O	C	CH2E	CH2P	CH2P									
Q	NH1	CH1E	O	C	CH2E	CH2E	C	O		NH2						
R	NH1	CH1E	O	C	CH2E	CH2E	CH2E	NH1	C	NC2	NC2					
S	NH1	CH1E	O	C	CH2E	OH1										
T	NH1	CH1E	O	C	CH1E	OH1				CH3E						
V	NH1	CH1E	O	C	CH1E	CH3E				CH3E						
W	NH1	CH1E	O	C	CH2E	C5W	CR1E			CW	NH1	CW	CR1E	CR1W	CR1E	CR1W
Y	NH1	CH1E	O	C	CH2E	CY	CR1E			CR1E	CR1E	CR1E	CY2	OH1		

Con esta información se puede ahora obtener de Engh y Huber (1991) las distancias de los enlaces covalentes y los ángulos de los enlaces que se necesitan como entrada para obtener las coordenadas de un átomo junto con un ángulo de torsión.

En las Tablas XIV y XV se muestran dos columnas para los átomos de la cadena lateral, una con encabezado Átomos Principales, y la otra Átomos Redundantes. A continuación se introduce el concepto de Átomos Principales y Átomos Redundates.

Los **átomos principales de la cadena lateral** son aquellos átomos pesados de la cadena lateral que se requieren para calcular los ángulos de torsión de la cadena lateral. Por ejemplo, en la valina se tiene un sólo ángulo de torsión para la cadena lateral, el cual se define por los átomos N, CA, CB y CG1, por lo tanto los átomos principales de la cadena lateral para la valina son CB y CG1.

En la Tabla XVI se muestra la relación de los átomos que definen a cada uno de los ángulos de torsión de la cadena lateral por cada aminoácido, los cuáles son los átomos principales de la cadena lateral.

Los **átomos redundantes de la cadena lateral** son aquellos átomos pesados de la cadena lateral que no se necesitan para el cálculo de los ángulos de torsión de la cadena lateral. Se les llama redundantes porque se pueden calcular a partir de los ángulos principales, ya que las propiedades químicas de los enlaces restringen el plano donde estos se encuentran. Usando el mismo ejemplo de la valina, se puede observar que además de los átomos principales (CB y CG1), la cadena lateral de la valina cuenta con un tercer átomo, CG2. Las coordenadas de este átomo se pueden calcular a partir de las coordenadas de los átomos principales además de las restricciones de distancia y ángulo de los enlaces covalentes.

Para que esto quede más claro, imagine que el i-ésimo residuo de una proteína es la valina. Usando una biblioteca de rotámeros se selecciona el rotámero $\mathbf{r} = (\chi_1)$. Para calcular las coordenadas del átomo CG1 se hace

Tabla XVI. Relación de átomos para determinar los ángulos de torsión de la cadena lateral para cada aminoácido.

Aminoácido		Ángulo	Átomos			
A	ALA	-	-	-	-	-
C	CYS	χ_1	N	CA	CB	SG
D	ASP	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	OD1
E	GLU	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD
		χ_3	CB	CG	CD	OE1
F	PHE	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD1
G	GLY	-	-	-	-	-
H	HIS	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	ND1
I	ILE	χ_1	N	CA	CB	CG1
		χ_2	CA	CB	CG1	CD1
K	LYS	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD
		χ_3	CB	CG	CD	CE
		χ_4	CG	CD	CE	NZ
L	LEU	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD1
M	MET	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	SD
		χ_3	CB	CG	SD	CE
N	ASN	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	OD1
P	PRO	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD
Q	GLN	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD
		χ_3	CB	CG	CD	OE1
R	ARG	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD
		χ_3	CB	CG	CD	NE
		χ_4	CG	CD	NE	CZ
		χ_5	CD	NE	CZ	NH1
S	SER	χ_1	N	CA	CB	OG
T	THR	χ_1	N	CA	CB	OG1
V	VAL	χ_1	N	CA	CB	CG1
W	TRP	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD1
Y	TYR	χ_1	N	CA	CB	CG
		χ_2	CA	CB	CG	CD1

$$
\begin{aligned}
\mathbf{A} &= \mathbf{n}_i \\
\mathbf{B} &= \mathbf{ca}_i \\
\mathbf{C} &= \mathbf{cb}_i \\
R &= 1.521^{1} \\
\theta &= 110.5^{o2} \\
\varphi &= \chi_1
\end{aligned}
$$

donde $\mathbf{n}_i$, $\mathbf{ca}_i$ y $\mathbf{cb}_i$ son las coordenadas de los átomos N, CA y CB, respectivamente, los cuales son datos de entrada. Con esto se resuelve $\mathbf{D}$ como se indica en la Sección A.1.5, y hacemos $\mathbf{cg1}_i = \mathbf{D}$.

Ahora bien, para obtener las coordenadas de CG2 no es necesario otro ángulo de torsión, sino que en base al mismo ángulo χ_1 se obtienen las otras coordenadas. Se hace,

$$
\begin{aligned}
\mathbf{A} &= \mathbf{n}_i \\
\mathbf{B} &= \mathbf{ca}_i \\
\mathbf{C} &= \mathbf{cb}_i \\
R &= 1.521^{1} \\
\theta &= 109.5^{o3} \\
\varphi &= \chi_1 + 120^{o3}
\end{aligned}
$$

y se sigue el mismo procedimiento antes mencionado.

[1]Distancia promedio de un enlace covalente entre un par de átomos tipo CH1E - CH3E (Engh y Huber, 1991).

[2]Ángulo promedio que forman los enlaces covalentes con los átomos tipo CH1E - CH1E - CH3E (Engh y Huber, 1991).

[3]Estos valores se muestran en la Tabla XVII.

Tabla XVII. Valores para generar las coodenadas de los átomos reduntandes de la cadena lateral.

Aminoácido	Átomo Redundante	**A**	**B**	**C**	φ	θ	R
ASP(D)	OD2	CA	CB	CG	χ_2+180	120	1.249
GLU(E)	OE2	CB	CG	CD	χ_3+180	120	1.249
PHE(F)	CD2	CA	CB	CG	χ_2+180	120	1.384
	CE1	CB	CG	CD1	180	120	1.382
	CE2	CB	CG	CD2	180	120	1.382
	CZ	CG	CD1	CE1	0	120	1.382
HIS(H)	CD2	CA	CB	CG	χ_2+180	130.15	1.355
	CE1	CB	CG	ND1	180	105.6	1.319
	NE2	CB	CG	CD2	180	106.85	1.374
ILE(I)	CG2	N	CA	CB	χ_1-120	109.5	1.521
LEU(L)	CD2	CA	CB	CG	χ_1+120	109.5	1.521
ASN(N)	ND2	CA	CB	CG	χ_2+180	120	1.328
GLN(Q)	NE2	CB	CG	CD	χ_3+180	120	1.328
ARG(R)	NH1	CD	NE	CZ	180	120	1.326
	NH2	CD	NE	CZ	0	120	1.326
THR(T)	CG2	N	CA	CB	χ_1-120	109.5	1.521
VAL(V)	CG2	N	CA	CB	χ_1+120	109.5	1.521
TRP(W)	CD2	CA	CB	CG	χ_2+180	126.8	1.433
	NE1	CB	CG	CD1	180	110.2	1.374
	CE2	CB	CG	CD2	180	107.2	1.409
	CE3	CB	CG	CD2	0	133.9	1.398
	CZ2	CG	CD2	CE2	180	122.4	1.394
	CZ3	CG	CD2	CE3	180	118.6	1.382
	CH2	CD2	CE2	CZ2	0	117.5	1.368
TYR(Y)	CD2	CA	CB	CG	χ_2+180	120	1.389
	CE1	CB	CG	CD1	180	120	1.382
	CE2	CB	CG	CD2	180	120	1.382
	CZ	CG	CD1	CE1	0	120	1.378
	OH	CD1	CE1	CZ	180	120	1.376

A.2 Discusión y perspectivas

Existen muchos detalles que están implícitos en el PSCPP los cuales aumentan la complejidad en la implementación del algoritmo.

El PSCPP, reducido a un problema de optimización combinatoria se compone de tres aspectos principales: la biblioteca de rotámeros, la función objetivo y el método de búsqueda. Debido a que las heurísticas más exitosas para el problema se basan en la técnica de recocido simulado, se seleccionó éste como el método de búsqueda. Sin embargo, aún hay que enfrentarse al problema de la selección de la biblioteca de rotámeros y la función objetivo.

Como ya se dijo anteriormente, las bibliotecas de rotámeros se obtienen estadísticamente, y se están generando continuamente más y mejores bibliotecas, por lo tanto lo mejor sería seleccionar la biblioteca más reciente. Un problema es que algunas bibliotecas no están disponibles, además, tienen diferentes formatos y no hay un estándar para éstas, lo cual implica implementar los métodos de lectura para cada biblioteca que se desea utilizar.

En este proyecto se utiliza la biblioteca independiente de la columna vertebral (Dunbrack y Cohen, 1997), pero cabe recalcar que hay bibliotecas más actualizadas (Dunbrack, 2002) que pudieran generar mejores resultados. La selección de la biblioteca se hizo para fines didácticos.

La selección de la función de energía también representa todo un reto, ya que los métodos actuales utilizan diversas funciones que pueden o no coincidir en algunos términos. Para el presente trabajo se está utilizando una parte de la función de energía utilizada en Bower *et al.* (1997). Ésta representa las interacciones del potencial de Lennard-Jones, y penaliza los choques interatómicos.

El algoritmo aquí descrito se pretende utilizar como base para uno más completo, que utilice una biblioteca de rotámeros más actualizada (una dependiente de la columna vertebral), que incorpore más términos en la función de energía, como las probabilidades de los rotámeros, para que pueda competir con los métodos actuales.