

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS



MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA
ÁREA DE NANOTECNOLOGÍA

UNA PROPUESTA DE BIBLIOTECA DE ROTÁMEROS DE PARES
DE RESIDUOS COMO APOYO A LOS MÉTODOS PARA EL
EMPAQUETAMIENTO DE LA CADENA LATERAL EN PROTEÍNAS

TESIS

que para cubrir los requisitos necesarios para obtener el grado de

DOCTOR EN CIENCIAS

Presenta:

SARA EUGENIA HERNÁNDEZ AYÓN

Ensenada, Baja California, México, julio de 2015

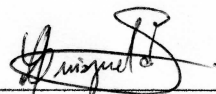
TESIS DEFENDIDA POR

Sara Eugenia Hernández Ayón

Y APROBADA POR EL SIGUIENTE COMITÉ



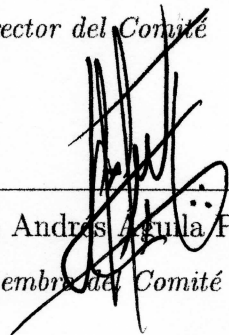
Dr. Jorge Alberto Villavicencio Aguilar
Codirector del Comité



Dr. Carlos Alberto Brizuela Rodríguez
Codirector del Comité



Dr. Israel Marck Martínez Pérez
Miembro del Comité



Dr. Sergio Andrés Aguila Puentes
Miembro del Comité



Dr. Faustino Camarena Rosales
Miembro del Comité

31 de julio de 2015

RESUMEN de la tesis de **SARA EUGENIA HERNÁNDEZ AYÓN**, presentada como requisito para la obtención del grado de DOCTOR EN CIENCIAS. Ensenada, Baja California, julio de 2015.

UNA PROPUESTA DE BIBLIOTECA DE ROTÁMEROS DE PARES DE RESIDUOS COMO APOYO A LOS MÉTODOS PARA EL EMPAQUETAMIENTO DE LA CADENA LATERAL EN PROTEÍNAS

Resumen aprobado por:

Dr. Jorge Alberto Villavicencio Aguilar

Codirector de Tesis

Dr. Carlos Alberto Brizuela Rodríguez

Codirector de Tesis

Predecir con alta precisión el empaquetamiento de las cadenas laterales en proteínas ha sido reconocido como un reto difícil. Los métodos que abordan este problema tienen tres elementos básicos: la biblioteca de rotámeros, la función de la energía, y el algoritmo de búsqueda. Las bibliotecas de rotámeros se han desarrollado sobre la base de un análisis estadístico de las conformaciones de las cadenas laterales de estructuras de proteínas conocidas. Estas bibliotecas son cada vez mejores con respecto a la cobertura de los ángulos de torsión, a expensas de ser cada vez más grandes y complejas. Bibliotecas con mayor tamaño tienen impacto en el espacio de búsqueda del algoritmo. El tamaño del espacio de búsqueda del algoritmo está acotado por $O(a^l)$ siendo l la longitud de la proteína, y a el número máximo de rotámeros por aminoácido. Es evidente que a medida que l aumenta, el tamaño del espacio de búsqueda aumentará exponencialmente. Motivados por esta observación, se propone un método para el diseño de bibliotecas de rotámeros que considera un par de aminoácidos consecutivos al mismo tiempo y que, en consecuencia, reduce el exponente del tamaño del espacio de búsqueda de l a $l/2$. El método utiliza una técnica sencilla de agrupación para generar la biblioteca conocida como k-modas. El método propuesto funciona de la siguiente manera. Dado un conjunto de estructuras de proteínas objetivo y un valor del nivel de cobertura de los ángulos de torsión deseado, para cada par de residuos contiguos en el conjunto de estructuras objetivo, se calcula el conjunto de rotámeros necesarios para alcanzar el nivel de cobertura especificada. Como prueba de concepto del método propuesto, se utilizó un conjunto de 149 estructuras de proteínas para generar la biblioteca. Los porcentajes de precisión alcanzable, calculados sobre dos conjuntos de prueba de 65 y 373 estructuras son: 96.7 %, 85.5 %, 78.9 % y 75.8 % para el primer conjunto y 96.2 %, 82.7 %, 73.5 % y 69.3 % para el segundo conjunto, sobre las medidas de los ángulos de torsión χ_1 , χ_{1+2} , χ_{1+2+3} y $\chi_{1+2+3+4}$, respectivamente.

Palabras Clave: Biología computacional, Predicción de estructura de proteínas, Empaquetamiento de la cadena lateral en proteínas, Bibliotecas de rotámeros, k-modas.

ABSTRACT of the thesis presented by **SARA EUGENIA HERNÁNDEZ AYÓN**, in fulfillment of the requirements of the degree of PH.D. IN SCIENCES. Ensenada, Baja California, July 2015.

A PROPOSAL OF ROTAMER LIBRARY OF PAIRS OF RESIDUES AS SUPPORT FOR PROTEIN SIDECCHAIN PACKING METHODS

Approved by:

Dr. Jorge Alberto Villavicencio Aguilar

Thesis Advisor

Dr. Carlos Alberto Brizuela Rodríguez

Thesis Advisor

The design of high accuracy predictors for the packing of side chains in proteins has been recognized as a tough challenge. Methods that tackle this problem have three basic elements: the library of rotamers, the energy function, and the search algorithm. Rotamer libraries have been developed based on statistical analysis of side chains conformations in known, high quality protein structures. These libraries are becoming better with respect to torsion angles coverage, at the expense of becoming larger and more complex. Larger libraries impact on the algorithms search space size. The size of the algorithms search space is bounded by $O(a^l)$ with l the protein length, and a the maximum number of rotamers among all amino acids. Therefore, as l increases the search space size will exponentially increase. Motivated by this observation, a method for designing rotamer libraries which consider a pair of consecutive amino acids at the same time and that, consequently reduces the exponent of the search spaces size from l to $l/2$, is proposed. The method uses a simple clustering technique known as k-modes to generate the library. The proposed method works as follows, given a set of target protein structures, and a value for a desired coverage level of torsion angles, for each pair of contiguous residues in the set of target structures, computes the set of rotamers needed to achieve the specified coverage level. As a proof of concept of the proposed method, a set of 149 protein structures was used to generate the library. The achievable accuracy percentages calculated on two test sets of 65 and 373 protein structures are: 96.7 %, 85.5 %, 78.9 %, and 75.8 % for the first test set, and 96.2 %, 82.7 %, 73.5 %, and 69.3 % for the second test set in torsion angles χ_1 , χ_{1+2} , χ_{1+2+3} , and $\chi_{1+2+3+4}$, respectively.

Keywords: Computational biology, Prediction of proteins structure, Packaging of the proteins side chain, Rotamer libraries, k-modes.

A mi familia

Agradecimientos

Gracias a Dios por darme la fuerza y la capacidad de entender y aprender para concluir este trabajo de tesis.

Con todo mi amor y cariño a mi esposo Gabriel Humberto por creer en mi capacidad, y brindarme su comprensión, cariño y amor. A mis hijos Sara Gabriela y Sergio Gabriel por ser mi fuente de motivación e inspiración. A mi mamá y hermanos quienes con sus palabras de aliento me exhortaron para que siguiera adelante y siempre fuera perseverante y cumpliera con mis ideales. A mi suegra por su apoyo y palabras de aliento brindadas.

Al Dr. Carlos Alberto Brizuela Rodríguez por su valiosa dirección y apoyo para seguir adelante hasta concluir este trabajo de tesis, por su asesoramiento experto, enseñanzas y comprensión. Al Dr. Jorge Alberto Villavicencio Aguilar, por todo su apoyo, tiempo, asesoramiento y experiencias durante este trabajo de tesis. Al Dr. Alejandro Martínez por despertar en mí el interés por el área de Nanotecnología.

A los miembros del Comité de tesis Dr. Israel Marck Martínez Pérez, Dr. Sergio Andrés Águila Puentes, y Dr. Faustino Camarena Rosales, por su tiempo, observaciones y sugerencias en el desarrollo de este trabajo.

A mis compañeros y amigos presentes y pasados, quienes sin esperar nada a cambio compartieron su conocimiento, alegrías y tristezas. Y a todas aquellas personas que durante estos años estuvieron a mi lado apoyándome y que lograron que este sueño se hiciera realidad.

A la UABC por permitirme estudiar este posgrado, y a su personal administrativo por la excelente atención que siempre me brindaron. Al CICESE por su hospitalidad durante esta estancia.

Al CONACyT por su apoyo económico para realizar mis estudios.

A todos Ustedes mi mayor reconocimiento y gratitud

Contenido

	Página
Resumen en español	I
Resumen en inglés	II
Dedicatoria	III
Agradecimientos	IV
Contenido	VI
Lista de Figuras	IX
Lista de Tablas	XI
I.. Introducción	1
I.1. Antecedentes y motivación	1
I.2. Definición del problema	5
I.3. Objetivos de la investigación	7
I.3.1. Objetivo general	7
I.3.2. Objetivos específicos	7
I.4. Metodología de la solución propuesta	7
I.5. Organización de la tesis	9
II.. Conceptos básicos	11
II.1. Proteínas	11
II.2. Aminoácidos	12
II.3. Estructura de las proteínas	13
II.3.1. Cadena principal	13
II.3.2. Cadena lateral	14
II.4. Ángulo de torsión	15
II.5. Rotámero	17
II.6. Problema de empaquetamiento de cadenas laterales de proteínas	20
II.6.1. Elementos principales de los métodos de PSCPP	21
II.6.2. Métodos del estado del arte para el PSCPP	21
III.. Biblioteca de rotámeros	23
III.1. Tipos de bibliotecas de rotámeros	24
III.1.1. Biblioteca de rotámeros dependiente de la cadena principal	25

Contenido (continuación)

	Página
III.1.2. Biblioteca de rotámeros dependiente de la estructura secundaria	25
III.1.3. Biblioteca de rotámeros independiente de la cadena principal	26
III.2. Análisis de las bibliotecas de rotámeros independientes de la cadena principal	29
IV.. Definición del problema del diseño de bibliotecas	33
IV.1. Formalización del problema	34
IV.1.1. Selección del conjunto de datos	35
IV.1.2. Criterios estadísticos para la generación de rotámeros	36
IV.2. Trabajo previo	38
IV.2.1. Biblioteca de Dunbrack	38
IV.2.2. Biblioteca de Lovell	39
V.. Metodología	41
V.1. Conjunto de datos seleccionados CD149	41
V.2. Biblioteca estándar (s-biblioteca)	42
V.3. Biblioteca de pares: enfoque naïf (p(s)-biblioteca)	43
V.4. Biblioteca de pares: un enfoque basado en agrupamientos (p-biblioteca)	44
V.4.1. Algoritmo principal	45
V.4.2. Cobertura de las bibliotecas	48
V.4.3. Correlación canónica	52
VI.. Resultados	54
VI.1. Conjuntos de prueba CP65 y CP373	54
VI.2. Precisión alcanzable de p-biblioteca sobre CP65 y CP373	55
VI.2.1. Precisión reportada	61
VII..Conclusiones y trabajo futuro	66
VII.1.Sumario	66
VII.2.Conclusiones	67
VII.3.Propuesta de trabajo futuro	68
REFERENCIAS	70
A.. Conjunto de datos CD149	74
A.1. Lista de estructuras en CD149	74

Contenido (continuación)

	Página
B.. Conjunto de prueba CP65	77
B.1. Lista de estructuras en CP65	77
C.. Conjunto de prueba CP373	80
C.1. Lista de estructuras en CP373	80
D.. Misceláneos	99

Lista de Figuras

Figura	Página
1. Convergencia de la Ciencia y la Tecnología.	2
2. Ejemplo de empaquetamiento de cadena lateral (proteína 1VJS, residuos Trp411, Thr412 y Arg413).	4
3. Metodología general para la generación de la biblioteca de pares de rotámeros.	8
4. Dogma central de la biología molecular. Inspirado en CienciaExplicada (2015).	12
5. Estructura general de los aminoácidos. Tomado de UNC (2006a). . . .	12
6. Los 20 aminoácidos que se presentan en las proteínas. Inspirado en Brändén y Tooze (1999).	14
7. Cadena principal en las proteínas. Tomado de UNC (2006b).	15
8. Átomos de la cadena latera para la Lisina y la Leucina. Inspirado en las gráficas de Pymol DeLano y Warren (2002).	16
9. Diagrama esquemático del concepto de ángulo de torsión. Tomado de Corona de la Fuente (2010).	17
10. Ángulos de torsión de la cadena principal del polipéptido Gly-Ala-Gly. Inspirado en Molecularsciences (2015).	17
11. Ángulos de torsión de la cadena lateral de la Lisina. Inspirado en CNX (2015).	18
12. Diagrama del método de cristalografía de rayos X. Tomado de CSIC (2015).	20
13. Diagrama esquemático para el diseño de bibliotecas.	34
14. Algoritmo para generar los rotámeros de pares.	46
15. Ejemplo de la distribución de los ángulos de torsión y de los rotámeros en p-biblioteca.	47
16. Algoritmo k -modas.	49
17. Rotámeros de pares de residuos en p-biblioteca y pares de residuos en CD149.	50
18. Número de rotámeros por pares de residuos en p-biblioteca.	50

Lista de Figuras (continuación)

Figura	Página
19. Número de rotámeros vs niveles de cobertura.	51
20. Comparación de tamaño de bibliotecas.	52
21. Correlación canónica de los ángulos de torsión del conjunto de datos CD149.	53
22. Abundancia de pares consecutivos de residuos en CP65 y CP373 orde- nado por residuo inicial.	55
23. Porcentajes de precisión alcanzable por residuo inicial de p-biblioteca en CP65.	57
24. Porcentajes de precisión alcanzable por residuo inicial de p-biblioteca en CP373.	58
25. Porcentaje de precisión global de la p-biblioteca probado en CP149, TS65 y TS373.	59
26. Porcentaje de precisión alcanzable global de la p-biblioteca en CP65 por grupo.	60
27. Porcentaje de precisión alcanzable global de la p-biblioteca en CP373 por grupo.	61

Lista de Tablas

Tabla	Página
I. Ángulos de torsión de las cadenas principal y lateral de los residuos. . .	19
II. Información básica de una biblioteca de rotámeros dependiente de la cadena principal.	26
III. Información básica de una biblioteca de rotámeros dependiente de la estructura secundaria.	27
IV. Información básica de una biblioteca de rotámeros independiente de la cadena principal.	28
V. Características de las principales bibliotecas de rotámeros independientes de la cadena principal.	29
VI. Cobertura alcanzable para las bibliotecas resultantes sobre el CD149. .	51
VII. Precisión alcanzable por diferentes bibliotecas.	62
VIII. Predicción de la cadena lateral mediante la selección del rotámero más probable.	63
IX. Precisión alcanzable de residuos enterrados vs. residuos expuestos. . . .	64
X. Residuos eliminados en CD149.	75
XI. Conjunto de prueba 65.	78
XII. Residuos eliminados en CP65.	79
XIII. Conjunto de prueba 373.	82
XIV. Residuos eliminados en CP373.	88
XV. Ángulos de torsión de la cadena lateral por tipo de residuo.	99
XVI. Porcentaje de precisión global de p-bibliotecas evaluadas en CP65 y CP373.	100

Capítulo I

Introducción

Este capítulo tiene como finalidad, brindar al lector una visión global de nuestro trabajo de tesis. La estructura del capítulo consta de cinco secciones: 1) Antecedentes y motivación, 2) Definición del problema; 3) Objetivos de la investigación; 4) Metodología de la solución propuesta; 5) Organización de la tesis.

I.1. Antecedentes y motivación

La característica fundamental de la nanotecnología es su carácter multidisciplinario ya que distintos campos de la ciencia y la tecnología (ver figura 1) convergen en ella. La ciencia y la tecnología han alcanzado un punto en el que las fronteras que separan a las diferentes disciplinas han empezado a diluirse. Es precisamente por esta razón que la nanotecnología se considera también una tecnología convergente que promete beneficios de todo tipo, desde nuevas aplicaciones médicas (De Silva, 2007) a soluciones de problemas ambientales (Martínez y Carbajal, 2012), por mencionar algunos, y muchos otros. Es precisamente en el área de aplicaciones médicas (nanomedicina) donde se trabaja en el desarrollo de nanotransportadores de fármacos a tejidos específicos del cuerpo, que pueden ser útiles en el tratamiento del cáncer u otras enfermedades; biosensores moleculares con la capacidad de detectar alguna sustancia de interés como glucosa o algún biomarcador de alguna enfermedad; nanorrobots programados para reconocer y destruir células tumorales o bien reparar algún tejido como el tejido óseo a raíz de una fractura; nanopartículas con propiedades antisépticas y desinfectantes, por mencionar algunos (Irache, 2008).

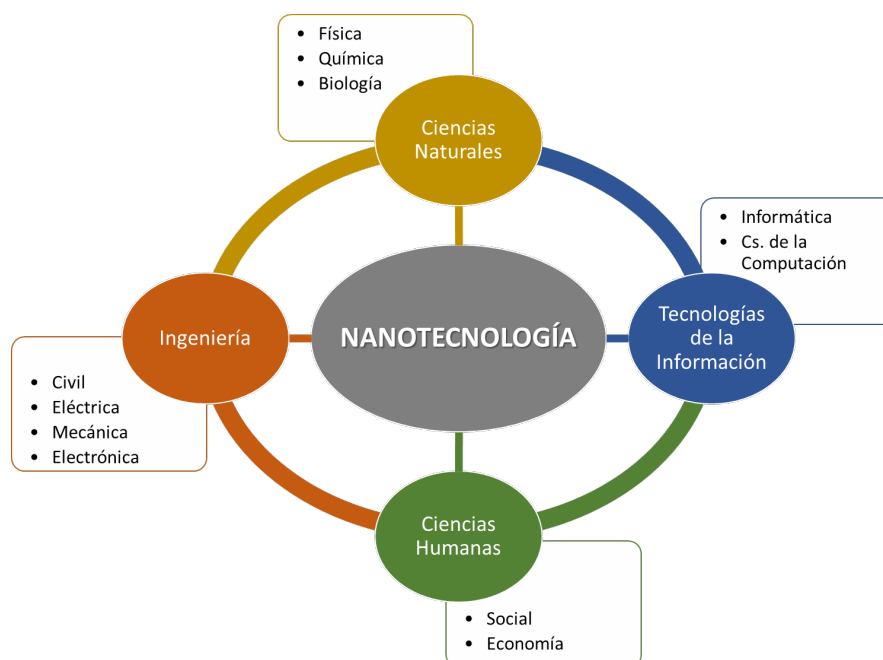


Figura 1. Convergencia de la Ciencia y la Tecnología.

En este contexto, un área de gran potencial en la nanomedicina es la ingeniería de proteínas, en donde se diseñan o modifican proteínas con la finalidad de habilitar su estructura para realizar una función específica. Las proteínas operan en células como eficientes máquinas programadas para modificar o producir moléculas (catalizadores) o funciones específicas. De ahí la importancia de estudiar sus estructuras tridimensionales complejas, las cuales están formadas por cientos o miles de átomos. En esta área se han desarrollado una serie de procedimientos orientados a la comprensión de la estructura de proteínas, así como al entendimiento de las relaciones estructura-función (Segovia y Peimbert, 2010). Dentro de la ingeniería de proteínas, uno de los problemas es la predicción de su estructura. La importancia radica en el hecho de que la función que realiza la proteína está relacionada con su estructura tridimensional, y al mismo tiempo, su estructura depende de la secuencia de aminoácidos que la conforman (Polaina, 2004). El proceso de obtención de la secuencia de una proteína es más sencillo que el de determinar su estructura. Actualmente, el número de secuencias de proteínas es

mayor a los 50 millones (UniProt, 2015), de las cuales menos de 110 mil tienen una estructura asociada (PDB, 2015). En la actualidad, la manera más confiable de obtener la estructura tridimensional de una proteína es a través de métodos experimentales de laboratorio, como lo son el método de cristalografía de rayos X y el método de resonancia magnética nuclear (Gu y Bourne, 2009). Sin embargo, obtener la estructura de manera experimental es un proceso que consume tiempo y dinero. Es por este motivo que surge la necesidad de diseñar métodos computacionales para predecir la estructura de proteínas a partir de su secuencia de aminoácidos (Sáenz *et al.*, 2007).

Existen tres métodos *in-silico** de predicción de estructuras de proteínas, los cuales son: modelado homólogo (Rodríguez *et al.*, 1998), reconocimiento de plegado (Bryant y Altschul, 1995) y *ab initio* (Osguthorpe, 2000). Este trabajo se enfoca al subproblema de modelado homólogo, denominado problema de empaquetamiento de la cadena lateral con proteínas (PSCPP, por sus siglas en inglés). Éste problema consiste en predecir las coordenadas tridimensionales de la proteína, a partir de la secuencia de aminoácidos que la conforman y las coordenadas de los átomos de la cadena principal de la misma. Akutsu en 1997, demostró que este problema es NP-Difícil, lo que implica que no se conoce algoritmo de tiempo polinomial alguno que lo pueda resolver y que la única alternativa exacta conocida requiere de la evaluación de un número exponencial de soluciones. En el área o contexto computacional, dicho problema de empaquetamiento se modela como uno de optimización combinatoria, en donde se recurre a una biblioteca de rotámeros que define el espacio de búsqueda y una función de energía que se debe minimizar (Krivov *et al.*, 2009). La biblioteca de rotámeros es una colección estadística de las distintas conformaciones que pueden tomar las cadenas laterales de los aminoácidos. Esta biblioteca se genera utilizando información de estructuras conocidas de proteínas. La función de energía a minimizar es una aproximación de la manera en la que los

* *in-silico*: hecho por computadora

átomos de la proteína interaccionan entre sí. Con base en esto, para solucionar este problema se desea encontrar la asignación de rotámeros, para cada uno de los residuos de la proteína de tal manera, que genere la menor energía posible.

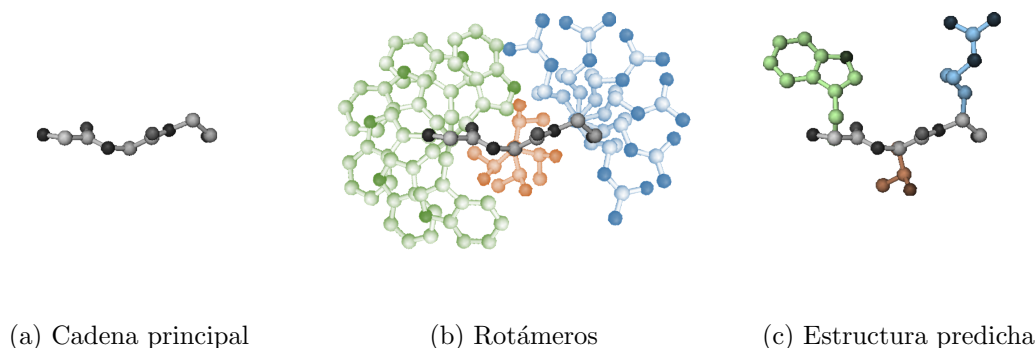


Figura 2. Ejemplo de empaquetamiento de cadena lateral (proteína 1VJS, residuos Trp411, Thr412 y Arg413).

En la figura 2 se muestra un ejemplo que ilustra el problema de empaquetamiento de la cadena lateral. En la figura 2 (a) se muestra la cadena principal de un segmento de tres residuos de la proteína 1VJS. En este caso la secuencia y las coordenadas de los átomos que conforman la cadena principal de los residuos Trp 411, Thr 412, y Arg 413 son considerados como datos de entrada. En la figura 2 (b) se muestra la cadena principal (en escala de grises) junto con las conformaciones de las cadenas laterales (rotámeros) de cada residuo (verde para Trp, rojo para Thr, y azul para Arg). En la figura 2 (c) se muestra la estructura predicha, conformada por la cadena principal y la cadena lateral para cada residuo. La energía de la estructura final mostrada en la figura 2 (c) debe ser menor a todas las estructuras posibles formadas por todas las combinaciones de los rotámeros mostrados en la figura 2 (b).

Como se definió anteriormente, un componente primordial del problema del empaquetamiento de la cadena lateral es la biblioteca de rotámeros. A continuación definimos de manera resumida el problema del diseño de la misma y se plantean los objetivos de

investigación de este trabajo.

I.2. Definición del problema

Existen diferentes tipos de bibliotecas de rotámeros, como lo son las dependientes de la cadena principal, las dependientes de la estructura secundaria y las independientes de la cadena principal. En este trabajo nos enfocaremos en las bibliotecas de rotámeros independientes de la cadena principal, cuyas principales representantes son: la de Dunbrack *et al.* (1997) y la de Lovell *et al.* (2000). El problema de diseño de bibliotecas de rotámeros independientes de la cadena principal, involucra diversos aspectos. El más importante es la generación de un grupo de rotámeros que se caracterizan por cubrir (en un porcentaje especificado) al conjunto de residuos (con sus respectivos ángulos de torsión) que lo originaron. En la literatura se reportan dos fases en el diseño de una biblioteca de rotámeros que a continuación describimos:

1. **Conjunto de datos.** En esta fase se realiza la selección, del conjunto de datos que dará origen a las bibliotecas. Para llevar a cabo la selección se establecen criterios como por ejemplo, que la resolución sea mejor que $1.8 \text{ \AA}$, o considerar un número de residuos por secuencia menor a 400, entre otros. Los tamaños de este conjunto de datos han sido de 19 (De Maeyer *et al.*, 1997), 30 (Liu *et al.*, 2003), 65 (Peterson *et al.*, 2004), 103 (Kono y Doi, 1996), 240 (Lovell *et al.*, 2000), y 518 (Dunbrack *et al.*, 1997) archivos PDB. En esta fase se considera además la incorporación de un filtrado del conjunto de datos que incluyen los hidrógenos, así como una verificación de las colisiones como en el caso de Lovell *et al.* (2000).
2. **Criterio estadístico.** Dentro de los criterios estadísticos para la generación de bibliotecas de rotámeros encontramos una gran variedad entre los que podemos

mencionar: la media en el trabajo de Kono y Doi (1996), la moda utilizada por Lovell *et al.* (2000), las estadísticas bayesianas de Dunbrack *et al.* (1997), y finalmente aquellas que fueron generadas con base en otras bibliotecas, como es el caso de De Maeyer *et al.* (1997); Liu *et al.* (2003); Peterson *et al.* (2004). Posteriormente ya sea se seleccionan u ordenan los rotámeros. Mediante este proceso se obtienen finalmente las bibliotecas, cuyos tamaños son de 153 (Lovell *et al.*, 2000), 315 (Dunbrack *et al.*, 1997), 335 (Kono y Doi, 1996), 549 (Liu *et al.*, 2003), 859 (De Maeyer *et al.*, 1997) y 49042 (Peterson *et al.*, 2004) rotámeros.

Se han desarrollado bibliotecas de rotámeros basadas en los análisis estadísticos de conformaciones de las cadenas laterales de estructuras de proteínas conocidas de alta calidad. Aún cuando las bibliotecas de rotámeros han mejorado con respecto de su cobertura de los ángulos de torsión, esto ha sido a costa de un incremento de su tamaño y complejidad. Ampliar las bibliotecas tiene un impacto en el tamaño del espacio de búsqueda del algoritmo. El tamaño de búsqueda del algoritmo está acotado por $O(b^n)$, siendo n la longitud de la proteína y b el número máximo de rotámeros entre todos los aminoácidos. Es decir, a medida que aumenta n , el espacio de búsqueda aumenta exponencialmente. Motivados por esta observación, en este trabajo proponemos un método para el diseño de bibliotecas de rotámeros. Este método considera un par de aminoácidos consecutivos al mismo tiempo y que, en consecuencia, reduce el exponente del tamaño del espacio de búsqueda de n a $n/2$.

I.3. Objetivos de la investigación

I.3.1. Objetivo general

El objetivo principal de nuestro trabajo de tesis es desarrollar una técnica para el diseño de bibliotecas de rotámeros de pares de residuos consecutivos, como apoyo a los métodos que solucionan el problema de empaquetamiento de las cadenas laterales en proteínas en la reducción del tamaño del espacio de búsqueda.

I.3.2. Objetivos específicos

- Caracterizar las propuestas de generación de bibliotecas de rotámeros existentes.
- Establecer criterios para la selección de proteínas que se utilizarán para la generación de la biblioteca.
- Diseñar una biblioteca de rotámeros individuales usando una técnica de agrupamiento.
- Diseñar una biblioteca de rotámeros de pares de residuos usando una técnica de agrupamiento.
- Determinar la calidad de las bibliotecas de rotámeros individuales y de pares propuestas.

I.4. Metodología de la solución propuesta

Para dar solución al problema anteriormente descrito, proponemos la creación de una biblioteca de pares de rotámeros. Para esto, se diseña una metodología que consiste de dos fases principales (Hernández y Brizuela, 2014), que a continuación describimos e ilustramos en la figura 3.

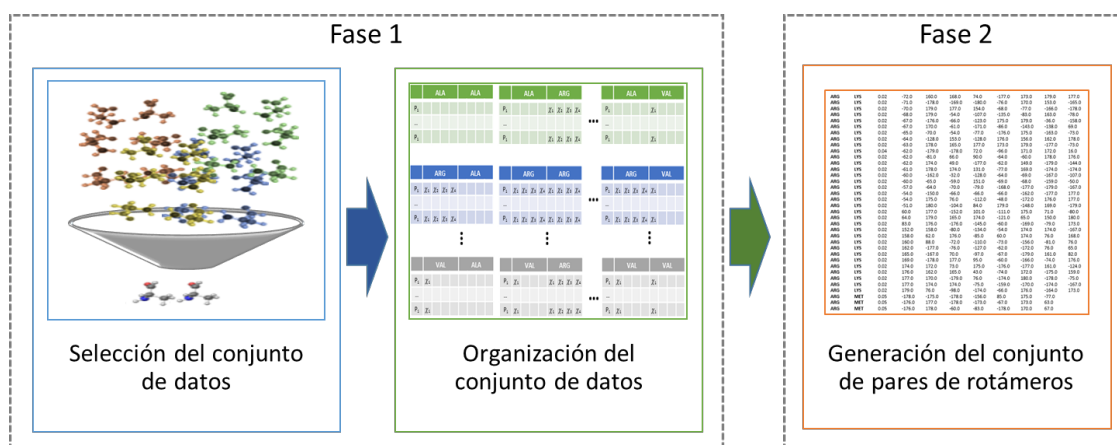


Figura 3. Metodología general para la generación de la biblioteca de pares de rotámeros.

1. **Selección del conjunto de datos.** Para generar la biblioteca se conforma un conjunto de datos que consiste de un grupo de estructuras de proteínas seleccionada con criterios específicos. Una vez obtenidas las estructuras en su formato PDB, se calculan los ángulos de torsión de las cadenas laterales de los residuos de cada una de estas estructuras. Del conjunto de datos se eliminan aquellos residuos cuyas cadenas laterales fueran incompletas; para la conformación de los pares de residuos, el criterio que se implementa es que estos sean consecutivos. Los pares de residuos se agrupan por tipo de pares de residuo, concatenando los ángulos de torsión de las cadenas laterales de los residuos que conforman las parejas.
2. **Generación del conjunto de pares de rotámeros.** Para generar los pares de rotámeros se toma cada uno de los grupos de pares de residuos con sus respectivos ángulos de torsión. Posteriormente, mediante el algoritmo conocido como k-modas (Huang, 1997) se obtienen los rotámeros correspondientes a cada par. Dichos pares de rotámeros se ordenan por residuo inicial, con lo que finalmente se obtiene la biblioteca de pares.

I.5. Organización de la tesis

El presente trabajo de tesis está organizado en siete capítulos y cuatro apéndices, cuyo contenido se describe de manera sucinta a continuación. En el capítulo 2 se presentan los conceptos básicos necesarios para la comprensión del problema tratado en este trabajo. En el capítulo 3 se presenta la descripción de los diversos tipos de bibliotecas de rotámeros, así como el análisis de las bibliotecas de rotámeros independientes de la cadena principal existentes. En el capítulo 4 se define a detalle el problema del diseño de bibliotecas. También se analiza el impacto que el concepto de pares de residuos tiene en el tamaño del espacio de búsqueda para los algoritmos de PSCPP. En el capítulo 5 se describe la metodología que seguimos para la generación de las bibliotecas de rotámeros individual y de pares. También se especifican los criterios de selección de las estructuras de proteínas que dieron origen a estas bibliotecas. Asimismo, se realiza una descripción del algoritmo propuesto como generador de bibliotecas, en donde además se incluye la información relacionada con la cobertura de estas bibliotecas sobre el conjunto de datos. Los resultados de la evaluación de la precisión (χ_1 , χ_{1+2} , χ_{1+2+3} , y $\chi_{1+2+3+4}$) de las bibliotecas individual y de pares sobre dos conjuntos de prueba, se presentan en el capítulo 6. En el capítulo 7 se exponen las conclusiones del presente trabajo de tesis, así como también algunas propuestas para la continuación de este trabajo de investigación. Con respecto del contenido de los apéndices del presente trabajo, en el apéndice A se listan las 149 estructuras proteicas que conforman el conjunto de datos. En este, también incluimos una relación de los residuos eliminados de dicho conjunto de datos por defectos en sus cadenas laterales. En los apéndices B y C se incluyen las listas de las proteínas que conforman los conjuntos de prueba (CP65 y CP373) utilizados en la evaluación de las bibliotecas. También se incluye la información acerca de los residuos eliminados de los conjuntos de prueba. Finalmente, en el apéndice D se incluye

información que profundiza algunos temas tratados en los capítulos principales.

Capítulo II

Conceptos básicos

En el presente capítulo se presentan los conceptos básicos necesarios para la mejor comprensión del problema tratado en este trabajo de tesis.

II.1. Proteínas

Una proteína es una macromolécula conformada por cadenas de aminoácidos plegadas en configuraciones particulares. Estas cadenas están unidas por enlaces peptídicos, por lo que a las proteínas también se les conoce como cadenas polipeptídicas. De acuerdo con el Dogma Central de la Biología Molecular (ver figura 4), las proteínas son el producto de los procesos conocidos como transcripción y traducción. En este proceso la información del ADN se transcribe a ARN, para posteriormente ser traducido a una secuencia de aminoácidos, finalmente esta cadena de aminoácidos se pliega produciendo una proteína. Este proceso es realizado en la célula en tiempos que van desde fracciones de segundo hasta unos pocos minutos (Chan y Dill, 1993).

Las proteínas son responsables de catalizar y regular reacciones bioquímicas (enzimas), como por ejemplo, las reductasas, hidrolasas y proteasas. Las proteínas también forman la base para componentes celulares y tejidos como piel, cabello y los tendones. Es conocido que la función de las proteínas está altamente relacionada con su estructura tridimensional (Brändén y Tooze, 1999) y que su estructura depende de la secuencia de aminoácidos que la compone.

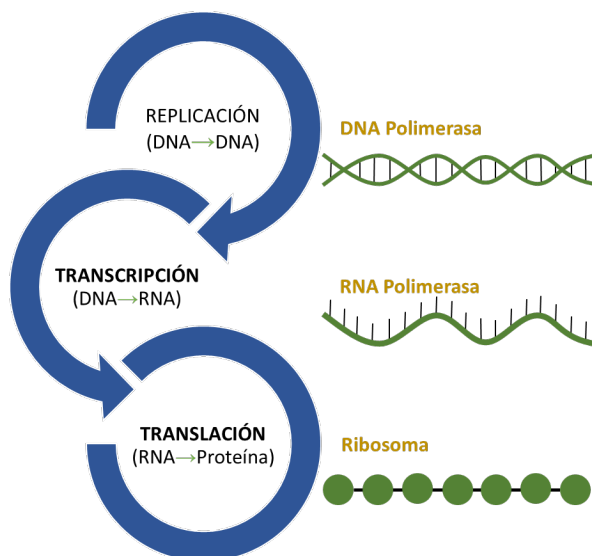


Figura 4. Dogma central de la biología molecular. Inspirado en CienciaExplicada (2015).

II.2. Aminoácidos

Se han encontrado 500 aminoácidos en la naturaleza (Wagner y Musso, 1983) de los cuales solo 20 son los más comúnmente utilizados para formar proteínas. Estos 20 aminoácidos cuentan con una estructura parecida entre ellos, la cual consiste de un grupo amino ($-NH_2$), un grupo carboxilo ($-COOH$) y un grupo R o cadena lateral unidos por un carbono conocido como C_α (ver figura 5).

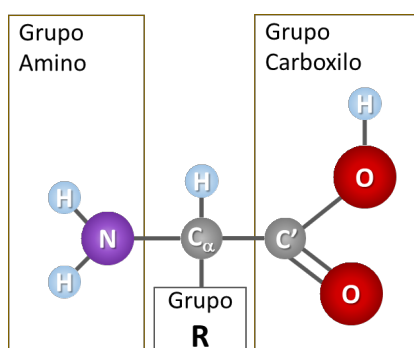


Figura 5. Estructura general de los aminoácidos. Tomado de UNC (2006a).

La mayoría de los aminoácidos existen en una de dos configuraciones denominadas estereoisómeros, que se designan con las letras D y L. Estas configuraciones son imágenes que corresponden a las formas tridimensionales “dextrógiro” (D) y “levógiro” (L). Los

aminoácidos encontrados en las proteínas siempre son L-isómeros (excepto la glicina, este aminoácido no tiene estereoisómeros). Los D-aminoácidos a veces aparecen en la naturaleza, en ciertas paredes celulares bacterianas y en antibióticos (Tortora *et al.*, 2007).

Las 20 cadenas laterales de los aminoácidos que se presentan en las proteínas se muestran en la figura 6. Sus nombres están abreviados con código de una y tres letras. Los aminoácidos son usualmente divididos en tres diferentes clases definidas por la naturaleza química de sus cadenas laterales. La primera clase comprende aquellos aminoácidos con cadena lateral estrictamente hidrofóbica: Ala(A), Val(V), Leu(L), Ile(I), Phe(F), Pro(P), y Met(M). Los cuatro aminoácidos cargados son: Asp(D), Glu(E), Lys(K), y Arg(R), que corresponden a la segunda clase. La tercera clase comprende aquellos aminoácidos con cadena lateral polar: Ser(S), Thr(T), Cys(C), Asn(N), Gln(Q), His(H), Tyr(Y), y Trp(W). La glicina (G), que tiene solo un átomo de hidrógeno como cadena lateral y es el aminoácido más simple de los 20, tiene propiedades especiales y usualmente es considerado o bien como una cuarta clase o como parte de la primera clase (Brändén y Tooze, 1999).

II.3. Estructura de las proteínas

II.3.1. Cadena principal

Los aminoácidos se unen extremo con extremo durante la síntesis de las proteínas para la formación de enlaces peptídicos cuando el grupo carboxilo ($COOH$) de un aminoácido condensa con el grupo amino (NH_2) del siguiente liberando una molécula de agua (figura 7). El proceso se repite varias veces hasta que se elonga la cadena. Como consecuencia de lo anterior, el grupo amino del primer aminoácido de la cadena polipeptídica y el grupo carboxilo del último aminoácido permanecen intactos y se dice

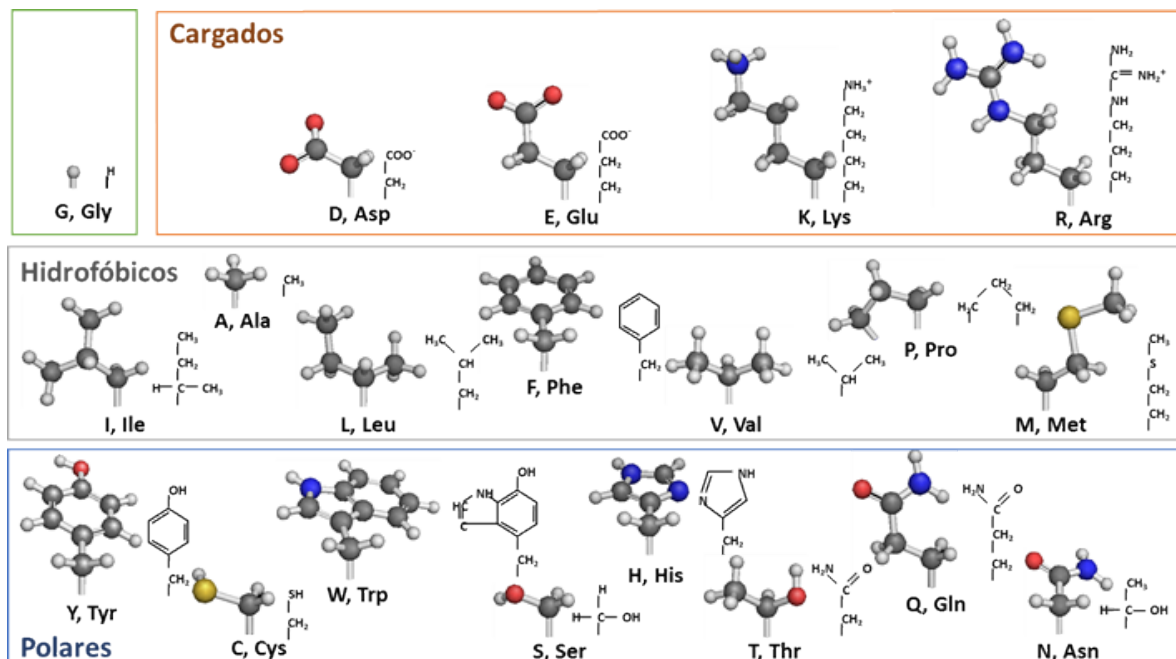


Figura 6. Los 20 aminoácidos que se presentan en las proteínas. Inspirado en Brändén y Tooze (1999).

que la cadena se extiende de su amino terminal a su carboxilo terminal. La formación de una sucesión de enlaces peptídicos genera una “cadena principal” (BB, backbone, por sus siglas en inglés), desde la cual se proyectan las diversas cadenas laterales.

Los átomos de la cadena principal corresponden a un átomo de carbono C_{α} al cual se le une la cadena lateral, un grupo NH, y un grupo $C'=O$. Estas unidades o residuos de aminoácidos son vinculadas en un polipéptido por un enlace peptídico entre el átomo C' de un residuo y el átomo de nitrógeno del siguiente residuo (figura 7). La unidad básica (residuo de aminoácido) que se repite a lo largo de la cadena principal es $NH-C_{\alpha}H-C'=O$.

II.3.2. Cadena lateral

La cadena lateral de un aminoácido es una secuencia de átomos, los cuales se etiquetan conforme al orden de separación con respecto al C_{α} , usando las letras griegas:

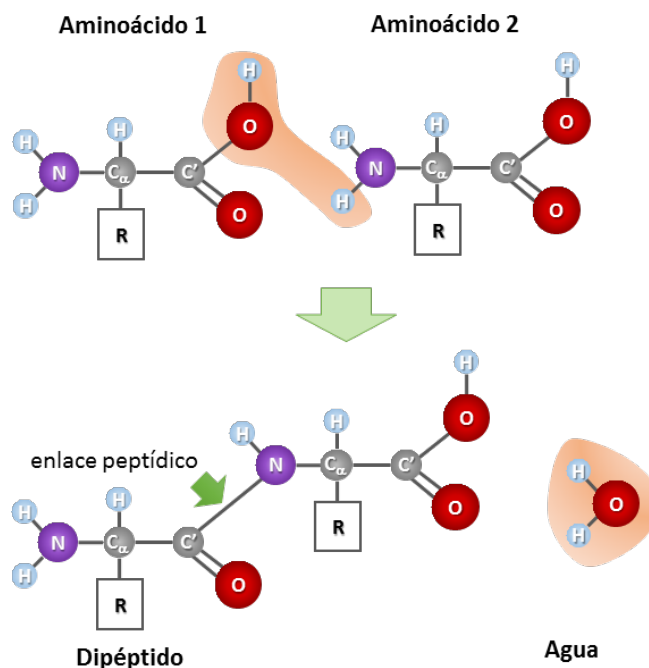


Figura 7. Cadena principal en las proteínas. Tomado de UNC (2006b).

$\beta, \gamma, \delta, \epsilon, \eta$, y τ , siendo el átomo etiquetado con β el más cercano al C_α . Además, si dos átomos se encuentran a la misma distancia de C_α , entonces se etiquetan con números arábigos (Creighton, 1993). En la figura 8 se muestra un ejemplo del etiquetado de los átomos de las cadenas laterales de la Lisina y la Leucina.

II.4. Ángulo de torsión

La rotación alrededor de los enlaces es descrita como función de los ángulos de torsión, que por lo general se encuentran en el intervalo de -180° y $+180^\circ$. Los valores de los ángulos de torsión se obtienen utilizando cuatro puntos en el espacio, (i, j, k, l) , en donde cada uno de estos puntos representa la posición de un átomo del aminoácido. Los puntos (i, j, k) definen el plano A y los puntos (j, k, l) el plano B . El ángulo formado entre los planos A y B (ver figura 9) se denota por φ_{ijkl} y puede calcularse mediante la siguiente fórmula:

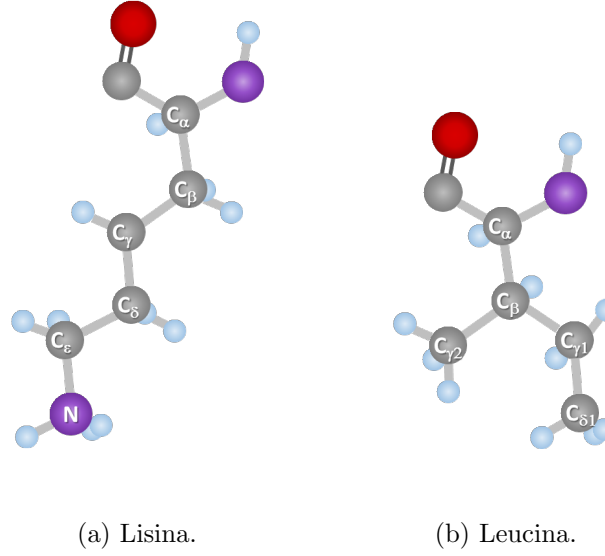


Figura 8. Átomos de la cadena lateral para la Lisina y la Leucina. Inspirado en las gráficas de Pymol DeLano y Warren (2002).

$$\varphi_{ijkl} = \cos^{-1}(\hat{n}_A \cdot \hat{n}_B), \quad (1)$$

donde $\hat{n}_A$ y $\hat{n}_B$ son los vectores normales unitarios de los planos A y B , respectivamente. Para calcular dichos vectores unitarios es necesario formar dos vectores para cada plano, para lo cual es necesario utilizar las coordenadas de los átomos que definen dicho plano.

Por ejemplo, en el caso del plano A , si cada átomo lo tratamos como un vector nulo, podemos crear los vectores $\vec{v}_0 = \vec{j} - \vec{i}$ y $\vec{v}_1 = \vec{k} - \vec{j}$; Con estos dos vectores podemos calcular la normal unitaria a A mediante $\hat{n}_A = v_0 \times v_1 / |v_0 \times v_1|$. Se sigue el mismo procedimiento para calcular $\hat{n}_B$, utilizando los vectores $\vec{v}_1 = \vec{k} - \vec{j}$ y $\vec{v}_2 = \vec{l} - \vec{k}$ (Rodriguez, 2014).

Los ángulos de torsión de un aminoácido se agrupan en dos categorías, los ángulos de torsión de la cadena principal, denominados ψ , ϕ y ω (ver figura 10) y los ángulos de torsión de la cadena lateral, nombrados como ángulos χ_i en donde i es un número

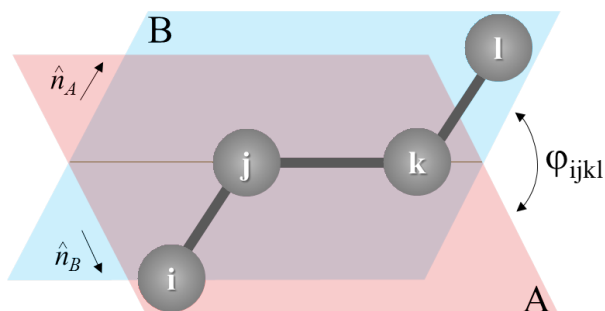


Figura 9. Diagrama esquemático del concepto de ángulo de torsión. Tomado de Corona de la Fuente (2010).

en orden ascendente que indica su cercanía al C_α (ver figura 11).

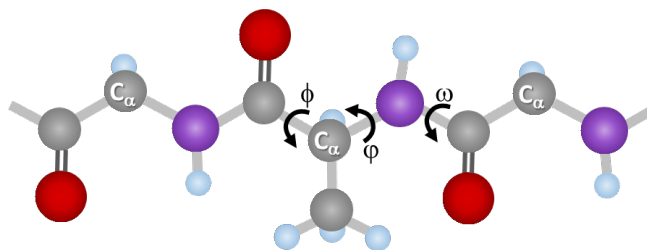


Figura 10. Ángulos de torsión de la cadena principal del polipéptido Gly-Ala-Gly. Inspirado en Molecularsciences (2015).

En la tabla I se muestran los ángulos de torsión junto con los cuatro átomos involucrados en su conformación. Cabe hacer mención que el número máximo de ángulos de torsión de la cadena lateral de los aminoácidos es de cinco, que es el caso de la Arginina. En contraste, al no contar con cadena lateral, la Alanina y la Glicina no poseen ángulos de torsión. En la tabla XV del apéndice D se encuentra la información completa de los ángulos de torsión de todos los residuos.

II.5. Rotámero

De acuerdo con un análisis de precisión para la determinación de estructuras de proteínas se ha demostrado que la mayoría de las cadenas laterales tienen una o pocas

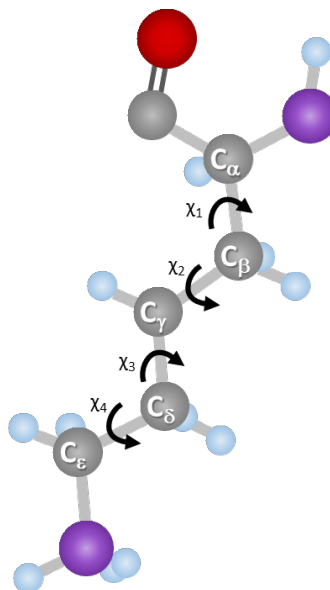


Figura 11. Ángulos de torsión de la cadena lateral de la Lisina. Inspirado en CNX (2015).

conformaciones que se producen con más frecuencia que otras, es decir, las conformaciones en las que la cadena lateral libera la energía de Gibbs mínima. Estas conformaciones son llamadas rotámeros (Brändén y Tooze, 1999) y se expresan como conjuntos de ángulos de torsión. Los rotámeros se obtienen utilizando proteínas con estructuras conocidas, es decir, proteínas que fueron sometidas a técnicas de laboratorio para obtener su estructura tridimensional, como lo son el método de cristalografía de rayos X y la técnica de resonancia magnética nuclear (Gu y Bourne, 2009).

La cristalografía de rayos X es la técnica más utilizada para determinar la estructura de proteínas y otras macromoléculas biológicas, el cual hasta abril de 2015 tiene registradas en PDB 96185 estructuras. El proceso que se sigue durante la cristalografía de rayos X para la determinación de la estructura en proteínas consiste en tomar una muestra purificada con altas concentraciones de una proteína y cristalizar dicha muestra. Después el monocristal se expone a rayos X, y con esto se producen más patrones de difracción, los cuales se analizan para poder producir así un mapa de la densidad de

Tabla I. Ángulos de torsión de las cadenas principal y lateral de los residuos.

Ángulo de torsión	átomos involucrados	cadena
ϕ	C - N - C_α - C	principal
ψ	N - C_α - C - N	principal
ω	C_α - C - N - C_α	principal
χ_1	N - C_α - C_β - $C_\gamma(1)$	lateral
χ_2	C_α - C_β - $C_\gamma(1)$ - $C_\delta(1)$	lateral
χ_3	C_β - $C_\gamma(1)$ - $C_\delta(1)$ - $C_\epsilon(1)$	lateral
χ_4	$C_\gamma(1)$ - $C_\delta(1)$ - $C_\epsilon(1)$ - $C_\zeta(1)$	lateral
χ_5	$C_\delta(1)$ - $C_\epsilon(1)$ - $C_\zeta(1)$ - $C_\eta(1)$	lateral

electrones sobre el cual se trata de encajar la estructura de la proteína de manera que adopte una conformación termodinámicamente favorable (ver figura 12).

En la técnica de resonancia magnética nuclear de protones (NMR) las propiedades magnéticas del espín de los núcleos atómicos dentro de una molécula se utilizan para obtener una lista de las restricciones de distancia entre los átomos de la molécula proteica, de la que se puede obtener una estructura tridimensional de la proteína. El método no requiere de proteínas cristalizadas y se puede utilizar en moléculas de proteínas en soluciones concentradas. Sin embargo, el método anterior está restringido en su uso a pequeñas proteínas (Brändén y Tooze, 1999).

Los rotámeros son almacenados en bibliotecas de rotámeros, las cuales son generadas estadísticamente, utilizando proteínas con estructura conocida y de ellas se obtienen los diferentes ángulos de torsión. A continuación se presenta uno de los problemas donde la biblioteca de rotámeros juega un papel fundamental.

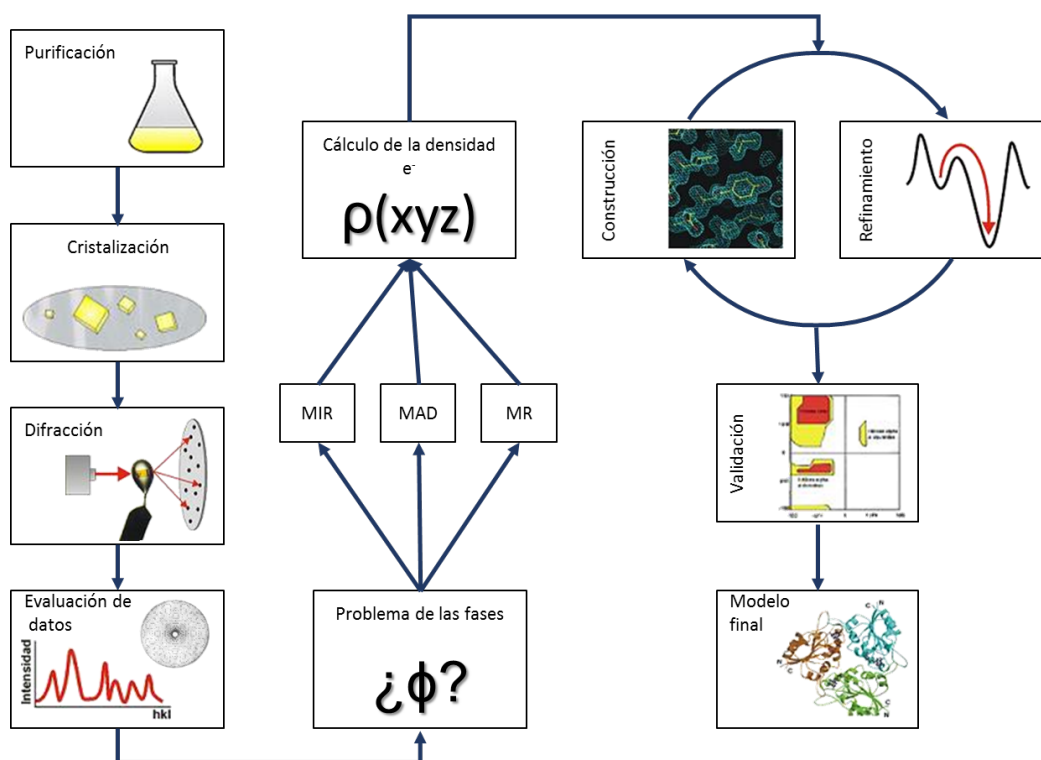


Figura 12. Diagrama del método de cristalografía de rayos X. Tomado de CSIC (2015).

II.6. Problema de empaquetamiento de cadenas laterales de proteínas

El problema de empaquetamiento de la cadena lateral de proteínas (PSCPP, por sus siglas en inglés) consiste en encontrar las coordenadas tridimensionales de cada uno de los átomos que conforman las cadenas laterales de una proteína conociendo la secuencia de aminoácidos y las coordenadas de los átomos de la cadena principal de la misma. Dentro de este problema es importante tener en cuenta que los átomos están sujetos a restricciones dadas por los enlaces entre átomos contiguos.

El PSCPP se presenta en dos problemas de bioinformática principalmente: 1) el modelado homólogo, que consiste en la predicción de la estructura tridimensional de secuencias de proteínas a partir de estructuras de secuencias homólogas (Venselaar *et al.*, 2009); 2) el rediseño de proteínas, en el que se requiere modificar parte de una estructura

conocida para que modifique sus propiedades, como aumentar su termoestabilidad, alterar su especificidad de enlace, mejorar su afinidad de enlace, aumentar su actividad enzimática, y alterar la especificidad al sustrato, entre otros (Gainza *et al.*, 2012).

II.6.1. Elementos principales de los métodos de PSCPP

En general los métodos para PSCP están compuestos por tres elementos principales (Miao *et al.*, 2011):

1. Una biblioteca de rotámeros de conformaciones de cadenas laterales. La mayoría de los métodos actuales dependen de rotámeros discretos, que puede reducir significativamente el costo computacional (Dunbrack *et al.*, 1997; Dunbrack y Karplus, 1993; Shapovalov y Dunbrack, 2011).
2. Una función de energía para la selección de los rotámeros. Se han propuesto un número de funciones de la energía para el empaquetado de la cadena lateral (Cao *et al.*, 2011; Krivov *et al.*, 2009; Lu *et al.*, 2008; Miao *et al.*, 2011).
3. Un algoritmo de búsqueda (Cao *et al.*, 2011; Krivov *et al.*, 2009; Lu *et al.*, 2008; Miao *et al.*, 2011) que busca la combinación entre rotámeros y la minimización de la función de energía.

II.6.2. Métodos del estado del arte para el PSCPP

A continuación se presentan un resumen de los principales métodos para la solución del problema de empaquetamiento de la cadena principal:

- OPUS-Rota (Lu *et al.*, 2008) implementa un nuevo término a la función de energía, el cual es un potencial estadístico para determinar la preferencia de orientación de las cadenas laterales. OPUS-Rota utiliza un algoritmo de recocido simulado por

baño de calor de Monte Carlo para llevar a cabo la optimización y la biblioteca de rotámeros dependiente de la cadena principal propuesta por Dunbrack y Karplus (1993).

- SCWRL4 (Krivov *et al.*, 2009) se basa en modelar las interacciones de la proteína como un grafo, el cual resuelve mediante algoritmos como el “Tree Decomposition” (Xu, 2005). Este algoritmo utiliza la biblioteca de rotámeros dependientes de la cadena principal desarrollada por Shapovalov y Dunbrack (2011).
- CIS-RR (Cao *et al.*, 2011) se enfoca en reducir el número de colisiones entre átomos al momento de predecir la estructura de la proteína. Una colisión entre una pareja de átomos ocurre cuando la distancia entre ambos es menor al 60 % de la suma de sus radios de van der Waals (Miao *et al.*, 2011). Este método utiliza la biblioteca de rotámeros dependiente de la cadena principal desarrollada por Dunbrack *et al.* (1997), en su versión del 2002.
- RASP (Miao *et al.*, 2011) de manera similar a CIS-RR, se enfoca en generar estructuras evitando las colisiones entre átomos. Sin embargo, este algoritmo primero genera una estructura de buena calidad utilizando distintos algoritmos, como por ejemplo de ramificación y terminación, búsqueda Monte Carlo y de marcha atrás. Este algoritmo utiliza la biblioteca de rotámeros dependiente de la cadena principal propuesta por Dunbrack *et al.* (1997).

Capítulo III

Biblioteca de rotámeros

Como se dijo en el capítulo anterior, la biblioteca de rotámeros es una discretización de los valores que pueden tomar los ángulos de torsión en una proteína, en donde se busca evitar que el algoritmo haga una búsqueda exhaustiva probando con todos los valores desde -180° a 180° , esperando proporcionar por medio de estas bibliotecas los valores más representativos que permitan cubrir la mayor parte de los ángulos de torsión.

Los rotámeros son almacenados en bibliotecas las cuales son generadas estadísticamente, utilizando los diferentes ángulos de torsión de proteínas con estructura conocida. Es por esto que cada rotámero de la biblioteca incluye la frecuencia y el valor de los ángulos de torsión asociados al rotámero, cuyo valor se encuentra en el intervalo de $[-180^\circ, 180^\circ]$. A través del tiempo se han publicado diferentes bibliotecas de rotámeros, esto es debido a que con el paso de los años se tienen disponibles más estructuras proteicas que se utilizan para generar mejores bibliotecas, así como el uso de diferentes métodos para la obtención de las mismas. Se ha demostrado (Lu *et al.*, 2008) que el generar bibliotecas de rotámeros muy grandes no necesariamente tiene como consecuencia mejores resultados. Esto se debe a que al momento de utilizar la biblioteca para la predicción de estructuras esto puede ocasionar que el método de búsqueda invierta más tiempo en encontrar las conformaciones que minimicen la energía libre. En contraste, el utilizar pocas estructuras proteicas puede generar una biblioteca específica que no tenga ángulos de torsión representativos de la población, por lo que ésta no se podría utilizar para la predicción de estructuras proteicas en general.

Existen varios tipos de bibliotecas, las cuales se describen en la siguiente sección.

III.1. Tipos de bibliotecas de rotámeros

Las bibliotecas de rotámeros pueden ser dependientes de la cadena principal (Dunbrack y Karplus, 1993; Dunbrack *et al.*, 1997; Shapovalov y Dunbrack, 2011), dependientes de la estructura secundaria (Lovell *et al.*, 2000; Schrauber *et al.*, 1993; McGregor *et al.*, 1987; Janin *et al.*, 1978), o independientes de la cadena principal (Dunbrack y Karplus, 1993; Dunbrack *et al.*, 1997; De Maeyer *et al.*, 1997; Liu *et al.*, 2003; Kono y Doi, 1996; Lovell *et al.*, 2000; Peterson *et al.*, 2004; Ponder y Richards, 1987; Schrauber *et al.*, 1993; Tuffery *et al.*, 1991). La clasificación se realiza como función de si los ángulos de torsión de los rotámeros dependen de la conformación de la cadena principal, de la estructura secundaria, o de ninguno, respectivamente.

En estos casos las bibliotecas dependientes de la cadena principal presentan conformaciones que son dependientes de la estructura local de la cadena principal, definidas por los ángulos de torsión ϕ y ψ (Dunbrack y Karplus, 1993; Dunbrack *et al.*, 1997; Shapovalov y Dunbrack, 2011). En las bibliotecas dependientes de la estructura secundaria los ángulos de torsión y/o las frecuencias de rotámeros dependen de si el aminoácido es parte de una estructura secundaria hélice- α , hoja- β , o giro (Lovell *et al.*, 2000; Schrauber *et al.*, 1993; McGregor *et al.*, 1987; Janin *et al.*, 1978). Las bibliotecas de rotámeros que son independientes de la cadena principal son calculadas a partir de todas las cadenas laterales disponibles de un determinado residuo (Dunbrack y Karplus, 1993; Dunbrack *et al.*, 1997; De Maeyer *et al.*, 1997; Liu *et al.*, 2003; Kono y Doi, 1996; Lovell *et al.*, 2000; Peterson *et al.*, 2004; Ponder y Richards, 1987; Schrauber *et al.*, 1993; Tuffery *et al.*, 1991). Recientemente, un nuevo tipo de biblioteca, conocido como biblioteca dependiente de la proteína, ha sido propuesto (Bhuyan y Gao, 2011). Esta

biblioteca es dependiente de la cadena principal y codifica información estructural de todos los residuos vecinos.

III.1.1. Biblioteca de rotámeros dependiente de la cadena principal

Las bibliotecas dependientes de la cadena principal contienen una colección de rotámeros, organizados por tipo de aminoácido para diferentes conformaciones de la cadena principal. Estos rotámeros cuentan con el conjunto de ángulos de torsión asociados a las cadenas laterales y se incluyen también los ángulos de la cadena principal (ψ y ϕ). Este tipo de bibliotecas son un poco más grandes, sin embargo, se obtienen mejores resultados con ellas (Dunbrack, 2002).

En la tabla II se muestra la información básica de algunos registros de una biblioteca de rotámeros dependiente de la cadena principal. A diferencia de la biblioteca de rotámeros independiente, esta biblioteca contiene la información de la conformación de la cadena principal, agrupando, a los residuos por tipo de aminoácido asociados a 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 (Arg) cuyos ángulos de torsión $\phi \bmod 10 = -180^\circ$ y $\psi \bmod 10 = -180^\circ$, otro conjunto para los residuos del aminoácido arginina cuyos ángulos de torsión $\phi \bmod 10 = -180^\circ$ y $\psi \bmod 10 = -170^\circ$, y así sucesivamente.

III.1.2. Biblioteca de rotámeros dependiente de la estructura secundaria

En las bibliotecas dependientes de la estructura secundaria podemos ver que la información está organizada con base en la clasificación de alfa hélices, láminas beta y giros.

Tabla II. Información básica de una biblioteca de rotámeros dependiente de la cadena principal.

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

En la tabla III se muestra la información básica de algunos registros de una biblioteca de rotámeros dependiente de la estructura secundaria. A diferencia de la biblioteca de rotámeros independiente, esta biblioteca incluye información de la conformación de la estructura secundaria, especificando un porcentaje a cada tipo de estructura secundaria α -hélice, lámina- β , y giros.

El uso de las bibliotecas dependientes de la estructura secundaria no es generalmente recomendable, ya que no proporcionan mucha más información que las bibliotecas de rotámeros independientes de la cadena principal (Dunbrack, 2002).

III.1.3. Biblioteca de rotámeros independiente de la cadena principal

Estas bibliotecas no hacen referencia a la conformación de la cadena principal y se calculan a partir de todas las cadenas laterales disponibles de cada uno de los residuos.

Tabla III. Información básica de una biblioteca de rotámeros dependiente de la estructura secundaria.

Tipo de Aminoácido	#	%	alfa	beta	otro	χ_1	χ_2	χ_3	χ_4
Arg	11	1 %	0 %	2 %	2 %	62°	180°	65°	−175°
Arg	16	2 %	1 %	2 %	2 %	62°	180°	180°	85°
Arg	15	2 %	1 %	2 %	2 %	62°	180°	180°	−85°
Arg	89	9 %	9 %	5 %	12 %	−67°	180°	180°	180°
...	...	...	...	...	...	...	...	...	...
Val	169	6 %	2 %	8 %	8 %	63°			
Val	1931	73 %	90 %	72 %	63 %	175°			
Val	526	20 %	7 %	20 %	28 %	−60°			

Estas bibliotecas proporcionan las probabilidades y conformaciones necesarias promediadas sobre las conformaciones observadas en la cadena principal de las proteínas.

En la tabla IV se presenta la información básica de algunos registros de una biblioteca de rotámeros independiente de la cadena lateral. Se sabe que para aminoácidos muy grandes como la arginina (Arg) se tienen hasta cinco ángulos de torsión en la cadena lateral, sin embargo, el ángulo χ_5 no se registra en las bibliotecas de rotámeros y al momento de utilizarse se establece como 180° (Corona de la Fuente, 2010). Para residuos pequeños como la valina (Val), que solo tiene un ángulo de torsión en la cadena lateral, el conjunto de rotámeros es menor. En este ejemplo solo se tienen tres rotámeros para la valina, para la cual, la suma de las probabilidades de éstos debe ser la unidad. Así mismo, la suma de todas las probabilidades de los rotámeros de la arginina debe ser uno, sin embargo, por facilidad aquí solo se muestran algunos registros. Cabe aclarar que tanto la alanina (Ala) como la glicina (Gly) no están consideradas en la biblioteca por no contar con cadena lateral. Estas bibliotecas son especialmente adecuadas para el refinamiento de estructuras de rayos X y NMR, ya que proporcionan

excelentes conformaciones iniciales que pueden ser alteradas cuando sea necesario por los datos experimentales. También son útiles para el desarrollo de escalas de entropía y para la representación de las propiedades del estado sin plegar de las estructuras proteicas (West y Smith, 1998). De acuerdo con Dunbrack (2002), la mejor biblioteca de este tipo es la de Lovell *et al.* (2000).

Tabla IV. Información básica de una biblioteca de rotámeros independiente de la cadena principal.

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

Recientemente ha surgido un nuevo tipo de biblioteca de rotámeros dependiente de la cadena principal, en donde se codifica la información estructural de todos los residuos vecinos. Esta biblioteca es conocida en la literatura como biblioteca de rotámeros dependiente de las proteínas (Bhuyan y Gao, 2011). En esta, dada una biblioteca de rotámeros y la estructura de la cadena principal de una proteína, primero se modela la estructura de la proteína como un campo aleatorio de Markov (Markov random field). Posteriormente, se estiman las distribuciones marginales con los algoritmos de inferencia, omitiendo el proceso de optimización o búsqueda global. Los rotámeros de la biblioteca dada son entonces reclasificados y asociados con las probabilidades actualizadas (Bhuyan y Gao, 2011).

III.2. Análisis de las bibliotecas de rotámeros independientes de la cadena principal

Dado que en este trabajo nos enfocamos en las bibliotecas de rotámeros independientes de la cadena principal, en la tabla V se muestra un resumen de las principales bibliotecas de este tipo. En esta tabla la primera columna muestra los autores que proponen la biblioteca, la segunda columna muestra el criterio estadístico para la obtención de los rotámeros, la tercera columna muestra el número de estructuras de proteínas analizadas por los autores para la obtención de la biblioteca, y finalmente, la cuarta columna muestra el número de rotámeros contenidos en la biblioteca.

Tabla V. Características de las principales bibliotecas de rotámeros independientes de la cadena principal.

Autor	Criterio para selección de rotámeros	Número de proteínas analizadas	Número de rotámeros en biblioteca
Kono y Doi (1996)	Promedio	103	335
De Maeyer <i>et al.</i> (1997)	Complementada con otras bibliotecas	19	859
Dunbrack <i>et al.</i> (1997)	Bayesiano	518	315
Lovell <i>et al.</i> (2000)	Moda	240	153
Liu <i>et al.</i> (2003)	Complementada con otras bibliotecas	30	525
Peterson <i>et al.</i> (2004)	Complementada con otras bibliotecas	65	49,042

- Kono and Doi proponen una biblioteca de rotámeros basada en la revisión de la biblioteca de Tuffery *et al.* (1991), utilizando 103 proteínas con un análisis de cluster estadístico. En esta biblioteca Arg (R) y Lys (K) fueron analizadas hasta el ángulo χ_4 , obteniendo 335 rotámeros en esta biblioteca, incluyendo frecuencias, promedios y varianzas de los ángulos diedros. Para generar los rotámeros, el método de cluster utiliza un algoritmo que comparte algunos procedimientos con el algoritmo k-media (Kono y Doi, 1996).
- La biblioteca de De Maeyer *et al.* (1997), fue generada con base en la biblioteca de Ponder y Richards (1987), completada con los rotámeros físicamente viables (conformaciones estándar gauche y trans). Dicha biblioteca es ampliada mediante la adopción de dos pasos de 10° alrededor del ángulo χ_1 en el caso de aminoácidos aromáticos (Phe, Tyr, His, y Trp), y para cada uno de estos nuevos rotámeros que dieron dos pasos de 20° alrededor del ángulo χ_2 , produciendo una biblioteca con 859 rotámeros (De Maeyer *et al.*, 1997).
- Dunbrack presenta un análisis estadístico Bayesiano de las cadenas laterales de las conformaciones de 850 proteínas del PDB. Se trata de una extensión de la biblioteca de rotámeros dependiente cadena principal que incluye las poblaciones de rotámeros y los promedios de ángulos de torsión para una amplia gama de valores de ϕ y ψ . Estas estadísticas Bayesianas requieren la suposición de una distribución *a priori* de los parámetros sobre su intervalo de valores posibles. Como resultado de lo anterior, se obtiene una biblioteca con 315 rotámeros, la cual incluye las probabilidades de rotámeros χ_2 , χ_3 , χ_4 , que están condicionadas por χ_1 (Dunbrack *et al.*, 1997).
- Siguiendo una serie de criterios predefinidos Lovell *et al.* (2000), elimina las cadenas laterales modeladas inadecuadamente. Es decir, muchas cadenas laterales

altamente tensas en las que probablemente sus ángulos de torsión sean promedios de rotámeros parcialmente ocupados y por lo tanto tienen ángulos de torsión inciertos. Los rotámeros también pueden ser removidos por una verificación de choque estérico. Estos autores utilizan la moda para definir los rotámeros, argumentando que las distribuciones de los ángulos de torsión pueden ser muy desiguales debido a las limitaciones físicas. La moda genera la conformación más común, mientras que la media puede dar una conformación parcialmente tensa. La biblioteca contaba originalmente 153 rotámeros y luego fue extendida hasta tener 173 rotámeros (Lovell *et al.*, 2000).

- La biblioteca de Liu *et al.* (2003), se construyó con base en la biblioteca de De Maeyer *et al.* (1997), así como con base en la combinación de obras de Tuffery *et al.* (1991), y Kono y Doi (1996). En este trabajo utilizan la desviación de corte de ángulos de torsión de la cadena lateral en 10° , 20° , 10° , y 10° para χ_1 , χ_2 , χ_3 , y χ_4 , respectivamente. Consideraciones subsiguientes toman Gly (G) y Ala (A) como si tuvieran una sola conformación, y Pro (P) como si tuviera dos conformaciones. También consideraron estructuras planares asimétricas tales como en His (H), Trp (W), Asp (D), y Gln (Q). De esta manera consiguieron una biblioteca de 549 rotámeros (Liu *et al.*, 2003).
- La biblioteca de Peterson *et al.* (2004), es la más grande hasta ahora descrita, con 49042 rotámeros. Esta biblioteca fue construida sobre los resultados del trabajo de Dunbrack *et al.* (1997). Se toman los ángulos de torsión y se mueven $\pm 15^\circ$ en pasos de 5° , obteniendo siete posiciones discretas para cada ángulo (+ g, -g, t) para un total de 21. También consideran altos grados de libertad para Arg (R), Lys (K), Glu (e), y Gln (Q); teniendo un tamaño de paso grueso de 15° para Lys (K) y Arg (R); y de 7.5° para Glu (E), Gln (Q) y Met (M), para 9 y 15 posiciones

discretas para cada diedro, respectivamente (Peterson *et al.*, 2004).

Conforme pasa el tiempo se ha observado que las bibliotecas de rotámeros son cada vez más complejas en su estructura y sus tamaños son cada vez más grandes. Sin embargo, las bibliotecas de gran tamaño constituyen un inconveniente para los algoritmos de búsqueda ya que el tamaño del espacio de búsqueda aumenta en orden $O(b^n)$, siendo n la longitud de secuencia de la proteína y b el número máximo de rotámeros de cualquier residuo. En este trabajo se considera que a pesar de su simplicidad las bibliotecas independientes de la cadena principal (s-biblioteca) son muy útiles. A pesar de que las bibliotecas de rotámeros para los residuos individuales proporcionan una buena precisión, el tamaño del espacio de búsqueda que inducen puede ser drásticamente reducido, si en lugar de trabajar con rotámeros para los residuos individuales, se consideran pares de residuos consecutivos.

Detallando un poco más, sean n la longitud de la secuencia de los residuos en una proteína; b_s el número promedio de rotámeros por aminoácido en la biblioteca rotámeros estándar (s-biblioteca); b_p el número promedio de rotámeros por pares de aminoácidos consecutivos en la biblioteca de rotámeros (p-biblioteca). Entonces, el tamaño del espacio de búsqueda para la s-biblioteca está acotado por $O(b_s^n)$, mientras que el espacio de búsqueda para p-biblioteca es $O(b_p^{n/2})$. Es claro que si $b_s = cb_p$ para una constante $c > 1$, entonces el tamaño del espacio de búsqueda para la p-biblioteca será exponencialmente más pequeño que uno inducido por la s-biblioteca. Esta reducción en el espacio de búsqueda es la principal motivación en este trabajo.

Capítulo IV

Definición del problema del diseño de bibliotecas

En el presente capítulo versa sobre la definición formal del problema del diseño de bibliotecas de rotámeros, individuales y de pares de residuos. Se presenta también una descripción de las bibliotecas de rotámeros independientes de la cadena principal más representativas.

Nuestro problema consiste en diseñar una biblioteca de rotámeros, con la peculiaridad de incorporar pares de residuos, en contraste con las bibliotecas hasta ahora conocidas que involucran solo residuos individuales (bibliotecas estándares). Con esta innovación se logra una reducción del tamaño del espacio de búsqueda siendo este resultado la principal motivación de nuestro trabajo de investigación. El problema toma como datos de entrada a un conjunto de estructuras de proteínas registradas en la base de datos de proteínas (alrededor de 110,000), a las cuales se aplican los criterios de selección para obtener el conjunto de estructuras con el que se pretende trabajar. De este conjunto resultante y usando criterios estadísticos (algoritmo k-modas) se genera un conjunto de rotámeros para cada tipo de aminoácido o por cada tipo de pares de aminoácidos, lo que permite obtener finalmente la biblioteca de rotámeros. En otras palabras, nuestro trabajo consiste en encontrar un conjunto de rotámeros de cardinalidad mínima, tal que este conjunto cubra un porcentaje específico de los ángulos de torsión de las estructuras de entrada.

A continuación se presenta en forma más rigurosa la definición del problema a resolver.

IV.1. Formalización del problema

A continuación se describen a detalle cada una de las fases del problema de diseño de bibliotecas de rotámeros, las cuales se presentan de forma esquemática en la figura 13.

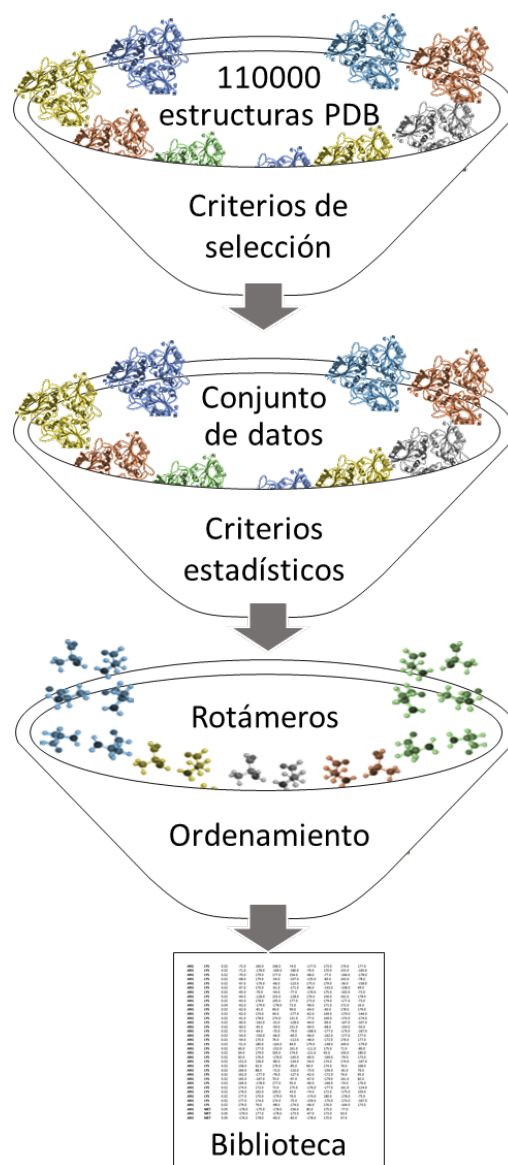


Figura 13. Diagrama esquemático para el diseño de bibliotecas.

IV.1.1. Selección del conjunto de datos

Para el diseño de bibliotecas de rotámeros es necesario seleccionar las estructuras de proteínas a partir de las cuales se construye la biblioteca. En este proceso se considera como población inicial de estructuras todas aquellas registradas en el banco de datos de estructuras de proteínas. Este banco de datos, mejor conocido como PDB, que cuenta a la fecha con alrededor de 110,000 estructuras (PDB, 2015). A partir de este conjunto de estructuras iniciales se establecen criterios y se seleccionan todas aquellas estructuras que satisfagan los criterios definidos. Esto permite obtener como resultado el conjunto de estructuras con el que se pretende trabajar. De acuerdo con la literatura, el tamaño de los conjuntos de estructuras seleccionados va desde 19 hasta 518 estructuras, como son los casos presentados por De Maeyer *et al.* (1997) y por Dunbrack *et al.* (1997), respectivamente. Así también dentro de esta fase, algunos autores realizan un procesamiento sobre las estructuras del conjunto de datos incorporando átomos de hidrógeno y verificando colisiones, como lo hace Lovell *et al.* (2000). De esta manera, se puede ser más exigente en el proceso de selección.

El subproblema de selección de las estructuras se define formalmente con las siguientes premisas.

Premisa. 1 Sea $E = \{e_1, e_2, \dots, e_N\}$ el conjunto de estructuras del PDB, donde N es el número total de estructuras, y e_i la estructura i -ésima. Cada e_i contiene una secuencia de aminoácidos $S_i = (a_1^i a_2^i \dots a_{|S_i|}^i)$, en donde a_j^i (aminoácido j de la estructura i) se le asocia un multiconjunto de ángulos de torsión $\tau_j^i = \{\psi_j^i, \phi_j^i, \omega_j^i, R_j^i\}$; aquí R_j^i representa el multiconjunto de ángulos de torsión de la cadena lateral del aminoácido a_j^i , tal que:

$$R_j^i = \{\chi_1^{i,j}\}, \text{ si } a_j^i \in \{Cys, Ser, Thr, Val\};$$

$$R_j^i = \{\chi_1^{i,j}, \chi_2^{i,j}\}, \text{ si } a_j^i \in \{Asp, Asn, His, Ile, Leu, Phe, Pro, Trp, Tyr\};$$

$$R_j^i = \{\chi_1^{i,j}, \chi_2^{i,j}, \chi_3^{i,j}\}, \text{ si } a_j^i \in \{Gln, Glu, Met\};$$

$$R_j^i = \{\chi_1^{i,j}, \chi_2^{i,j}, \chi_3^{i,j}, \chi_4^{i,j}\}, \text{ si } a_j^i \in \{Arg, Lys\}.$$

Premisa. 2 Sea $C = \{c_1, c_2, \dots, c_K\}$ un conjunto de criterios de selección, es decir, c_j puede ser una resolución mínima, o una cota máxima para el factor B , entre otras.

Dadas estas premisas el problema de selección de estructuras queda definido de la siguiente manera:

Dado E y un conjunto de criterios $C' \subseteq C$ seleccionar un conjunto de estructuras $P \subseteq E$ que satisfagan cada criterio $c_i \in C'$.

IV.1.2. Criterios estadísticos para la generación de rotámeros

Como resultado de la primera fase se obtienen los ángulos de torsión asociados a las estructuras seleccionadas. Posteriormente, se aplican a dichas estructuras diversos criterios estadísticos con distintos niveles de complejidad. Por ejemplo, en el trabajo de Kono y Doi (1996); identificamos el uso de la media; por su parte Lovell *et al.* (2000) utilizan la moda; Dunbrack *et al.* (1997), emplean estadísticas bayesianas. Finalmente los trabajos de De Maeyer *et al.* (1997), Liu *et al.* (2003) y Peterson *et al.* (2004), generan sus representantes (rotámeros) con base a la combinación de otras bibliotecas como son las de Ponder y Richards (1987), Tuffery *et al.* (1991), Kono y Doi (1996) y Dunbrack *et al.* (1997). Estos criterios estadísticos permiten elegir a los representantes que cubran al mayor número de estructuras del conjunto de datos.

Una formalización para el problema de seleccionar ángulos de torsión representativos (rotámeros) se plantea a continuación:

Premisa. 3 Sea A_X la colección de todos los ángulos de torsión de la cadena lateral correspondiente al aminoácido tipo X ($X = \{Ala, Arg, Asn, \dots, Val\}$) en el conjun-

to $P \subseteq E$ de estructuras de entrada. Por ejemplo, si $X = Cys$, entonces $A_{Cys} = \{\chi_1^{1,1}, \chi_1^{1,2}, \dots, \chi_1^{1,K_{Cys}}, \dots, \chi_1^{N,K_{Cys}}, \}$, donde (i, K_{Cys}) representa el número de cisteínas en la estructura i .

Premisa. 4 Sea también $A_{X/Z}$ la colección de todos los ángulos de torsión de la cadena lateral correspondiente al par consecutivo de aminoácidos X/Z ($X/Z = \{Ala/Ala, Ala/Arg, \dots, Val/Val\}$) en el conjunto de estructuras de entrada. Por ejemplo si $X = Cys$ (con un ángulo de torsión) y $Z = Leu$ (con dos ángulos de torsión), entonces $A_{Cys/Leu} = \{\chi_1^{1,1}(l)\chi_1^{1,1}(l+1)\chi_2^{1,1}(l+1), \dots, \chi_1^{1,P_{Cys/Leu}}(l)\chi_1^{1,P_{Cys/Leu}}(l+1)\chi_2^{1,P_{Cys/Leu}}(l+1), \dots, \chi_1^{N,P_{Cys/Leu}}(l)\chi_1^{N,P_{Cys/Leu}}(l+1)\chi_2^{N,P_{Cys/Leu}}(l+1)\}$, donde l y $l+1$ son los índices del aminoácido l y del $l+1$, respectivamente, y $(i, P_{Cys/Leu})$ es el número de pares consecutivos de Cys/Leu en la estructura i . En los conjuntos anteriores, χ_1 del residuo l -ésimo es el ángulo de torsión que corresponde a Cys y $\chi_1\chi_2$ son los ángulos de torsión de Leu en la posición $l+1$.

Premisa. 5 Sea $Cr = \{cr_1, cr_2, \dots, cr_K\}$ un conjunto de criterios, donde, cr_j puede representar el uso de media, moda, o estadística bayesiana, entre otros.

Dadas estas premisas el problema de selección de representantes (rotámeros) queda definido de la siguiente manera:

Dados A_X o $A_{X/Z}$ y un conjunto de criterios $C'r \subseteq Cr$ seleccionar un conjunto de representantes (rotámeros) $sBR \subseteq A_X$ para la biblioteca estándar o $pBR \subseteq A_{X/Z}$ para la biblioteca de pares que satisfagan cada criterio $cr_i \in C'r$.

Para propósitos de esta tesis, $C'r = \{cr_1, cr_2, cr_3\}$. En esta notación cr_1 representa el uso del algoritmo de agrupamiento k-modas, cr_2 la cobertura de los ángulos de torsión de P mayor o igual a un porcentaje especificado y cr_3 el conjunto de rotámeros de

cardinalidad mínima. Lo anterior aplica para ambas bibliotecas estándar y de pares de residuos. A continuación se toma cada rotámero generado previamente y se calcula su probabilidad acorde a su frecuencia. La probabilidad (frecuencia) se asigna de acuerdo al número del conjunto de datos que cada uno de los rotámeros que cubre. Como resultado final de este proceso se obtiene la biblioteca de rotámeros. La biblioteca de rotámeros, está conformada por un listado de rotámeros cuya estructura en general cuenta con los siguientes campos: identificación del residuo, la probabilidad del rotámero y los ángulos de torsión. Con respecto de estos últimos, se obtienen los ángulos $(\chi_1, \chi_2, \chi_3 \text{ y } \chi_4)$ en el caso de la biblioteca de residuos individuales o $(\chi_1, \chi_2, \chi_3, \chi_4 / \chi_1, \chi_2, \chi_3, \chi_4)$ en el caso de la biblioteca de rotámeros de pares de residuos. Aunque las fases son las mismas para todas las bibliotecas, el tamaño de estas varía, ya que como vimos en el capítulo anterior, el uso de diferentes conjuntos de datos y de criterios estadísticos conducen a estas variaciones de tamaño. Dichas variaciones van desde 153 rotámeros como es el caso de la biblioteca de Lovell *et al.* (2000), hasta 49042 como es el caso de la de Peterson *et al.* (2004).

IV.2. Trabajo previo

Dentro de las bibliotecas de rotámeros independientes de la cadena principal, las más representativas por su popularidad y eficiencia son la de Dunbrack y la de Lovell, las cuales describimos a continuación.

IV.2.1. Biblioteca de Dunbrack

Dunbrack *et al.* (1997), utilizan la estadística bayesiana para estimar las poblaciones y los ángulos de torsión para todos los rotámeros de todo tipo de cadena lateral en todos los valores de ϕ y ψ . Esto se logra mediante la derivación de una distribución

informativa *a priori* basada en el producto de las dependencias de ϕ y ψ . Utilizando el formalismo Bayesiano para combinar la probabilidad de datos totalmente dependientes de ϕ y ψ (una distribución multinomial) con las distribuciones previas expresadas como funciones de Dirichlet. En partes pobladas del mapa de Ramachandran, los resultados de este cálculo son las poblaciones y los ángulos de torsión muy cercanos a los valores de los datos; en las regiones poco pobladas del mapa de Ramachandran, la distribución previa informativa domina la población y ángulos predichos. El mecanismo bayesiano permite una transición suave entre estas dos situaciones (Dunbrack, 2002). En el enfoque de estos autores se considera a los ángulos de torsión en cada compartimento como si procedieran de distribuciones normales y utilizan un tratamiento bayesiano para estimar la varianza para cada ángulo de torsión. La varianza χ_1 se calcula como función de los ángulos de torsión de la cadena principal, mientras que las varianzas de χ_2 - χ_3 - χ_4 se consideran dependientes de χ_1 , pero independientes de la conformación de la cadena principal.

IV.2.2. Biblioteca de Lovell

Lovell *et al.* (2000), utilizan el criterio de eliminación de cadenas laterales incorrectamente modeladas. Esto es, muchas cadenas laterales tienen ángulos de torsión altamente forzados, probablemente como resultado de promedios de rotámeros parcialmente ocupados y por lo tanto cuentan con ángulos de torsión que no se explican. En el enfoque de estos autores, también puede remover cadenas laterales después de una verificación de choque estérico. Una vez hecha la eliminación, se utiliza la moda en lugar de la media para definir los rotámeros. Lo anterior se debe a que las distribuciones de los ángulos de torsión pueden ser muy sesgadas debido a restricciones físicas, como por ejemplo los autores mencionan que la moda genera la conformación más común, mientras que la media puede producir conformaciones parcialmente forzadas. Esto se

logra mediante el uso de una función de suavizado gaussiano en cada punto de datos y el cálculo de la densidad resultante de cada grado para identificar la moda. La variación de los ángulos de torsión estuvo representada por la anchura-media a la mitad del máximo de las distribuciones de Gauss suavizadas, después de corregir para la función de suavizado. Para distribuciones normales, estos valores son más grandes que la desviación estándar (σ) por un factor de 1.174. Sin embargo, dada la no normalidad de las distribuciones, Lovell *et al.* (2000), utilizan el promedio de los dos valores de la anchura media. También examinan el sesgo considerando diferentes anchuras medias en cada lado de la moda. Esta es una buena solución para el problema de la distribución de ángulos diedros. Porque es efectivamente un enfoque no paramétrico y no depende de supuestos de que los ángulos de torsión sean el resultado del azar, sino más bien se basa en una función de distribución bien definida. Sin embargo, hay un inconveniente potencial que ocurre cuando los rotámeros son relativamente raros, arrojan estimaciones pobres de la moda (Dunbrack, 2002).

Capítulo V

Metodología

En este capítulo se propone la metodología para la creación de una biblioteca de rotámeros de pares de residuos, que consistió de tres fases principales: 1) selección del conjunto de datos, 2) organización del conjunto de datos, y 3) generación del conjunto de rotámeros de pares. Para un mejor entendimiento, este capítulo primeramente describe el conjunto de datos seleccionado correspondiente a la primera fase de la metodología. Posteriormente se describen las bibliotecas estándar, de pares (con enfoque naïf) y al final se enfatiza en la biblioteca de rotámeros de pares de residuos (con enfoque basado en agrupamientos).

V.1. Conjunto de datos seleccionados CD149

El conjunto de datos para generar la biblioteca consiste en 149 estructuras de proteínas que llamamos CD149. Este fue obtenido de 240 estructuras de proteínas filtradas de Lovell *et al.* (2000). De acuerdo con Dunbrack (2002), esta es la mejor biblioteca de rotámeros independientes de la cadena principal porque incluye un procesamiento de las estructuras tales como la inclusión de átomos de hidrógeno y el removido de cadenas laterales no confiables. Además de estas 240 estructuras, 149 se mantienen en la lista Top500 de estructuras del Richardson Lab (Laboratory, 1999). El conjunto de datos resultante tiene las siguientes características: resolución de $1.8 \text{ \AA}$ o mejor, puntaje de colisiones* (para átomos con $\text{factor-}R < 40$) $< 22/100$ átomos (Word *et al.*, 1999), menos de $10/1000$ átomos cuyos principales ángulos de enlace de la cadena principal

* *puntaje de colisiones*: conteo del número total de colisiones entre cada par de átomos

son mayores de cinco desviaciones estándar de la geometría de Engh y Huber (Engh y Huber, 1991), sin sustitución de cadenas principales de aminoácidos inusuales, sin refinamiento de átomos libres, preferencia por tipo nativo a mutante, y si las proteínas están relacionadas, pero no son iguales, toma la mejor combinación de resolución y la puntuación de colisiones. Por lo tanto, el conjunto de datos consta de 149 estructuras de proteínas (véase la lista A.1 en apéndice A), es decir, 28543 aminoácidos de los cuales 226 fueron detectados con cadenas laterales incompletas, por lo que el montaje se reduce en 0.8 % quedándose con 28317 aminoácidos útiles. La lista de los identificadores PDB de las estructuras junto con una lista que muestra los aminoácidos eliminados están indicados en la tabla X del apéndice A.

V.2. Biblioteca estándar (s-biblioteca)

En este enfoque, para la organización del conjunto de datos, se conformaron los grupos por tipo de residuo. Para la generación del conjunto de rotámeros, utilizamos el algoritmo conocido como k-modas (Huang, 1997) obteniendo los rotámeros correspondientes a cada tipo de residuo. Los rotámeros fueron ordenados por su tipo de residuo para generar la biblioteca estándar (s-biblioteca).

En un trabajo reciente (Hernández y Brizuela, 2014) propusimos el uso del algoritmo k-medias para generar una biblioteca de rotámeros por aminoácido. Sin embargo, como lo señalaron Lovell *et al.* (2000), resulta más conveniente utilizar la moda que la media. Por un lado, esto se debe a que la distribución de los valores de los ángulos de torsión no es gaussiana, lo cual es una suposición implícita cuando se calcula la media y la desviación estándar. Por otro lado, tenemos que en el caso de la moda no es necesario hacer suposiciones *a priori* sobre el número y ubicación de los picos en la distribución. Además, se observó que si la distribución es convertida en una energía equivalente, es la

moda y no la media la que corresponde a la conformación de menor energía. Por estas razones se selecciona la moda como el criterio para agrupar los ángulos de torsión.

V.3. Biblioteca de pares: enfoque naïf (p(s)-biblioteca)

Con el objetivo de explorar la posibilidad de reducir el tamaño del espacio de búsqueda para el algoritmo, tomando pares de rotámeros para residuos consecutivos, consideramos todos los rotámeros de una biblioteca estándar (s-biblioteca), que fue generada a partir del conjunto de datos CD149 mediante la modificación de nuestro algoritmo k-means (Hernández y Brizuela, 2014), donde k-means se sustituyó por k-modas.

Desde la biblioteca estándar (s-biblioteca) obtenida, se generan todas las combinaciones posibles, lo que resulta en una matriz de dos dimensiones. A la biblioteca resultante la llamamos p(s)-biblioteca, es decir, una biblioteca de pares creados a partir de una biblioteca estándar. Por ejemplo, si solo tuvieramos dos residuos (A y C), A con dos rotámeros (A_1, A_2) y C con tres rotámeros (C_1, C_2, C_3) en la biblioteca estándar, para el par AC , obtendríamos seis rotámeros de pares de residuos consecutivos ($A_1-C_1, A_1-C_2, A_1-C_3, A_2-C_1, A_2-C_2$, y A_2-C_3). En la expresión anterior, A_i y C_j representan el rotámero i -ésimo de A y el rotámero j -ésimo de C , respectivamente. Es necesario realizar este tipo de combinaciones para cada par posible de la forma (CA, AA, CC) . Siguiendo esta idea obtenemos como resultado la p(s)-biblioteca con 134672 pares de rotámeros procedentes de una s-biblioteca con 368 rotámeros. En esta biblioteca el par de residuos con el mayor número de rotámeros es Lys-Lys con 14884 rotámeros de pares. Los intervalos para el número de rotámeros de todos los pares van de 0 (Ala-Ala, Ala-Gly, Gly-Ala, y Gly-Gly) a 14884 (Lys-Lys).

Ahora mostraremos que el construir una biblioteca de rotámeros de esta manera

resulta no ser una buena idea. Sea $O(b_S^n)$ el tamaño del espacio de búsqueda inducido por una s-biblioteca que tiene en promedio b_S rotámeros por aminoácido, y n es el número de residuos en la proteína. Mediante la generación de rotámeros de pares que resultan de la combinación de los rotámeros de la s-biblioteca todos contra todos, obtendremos un tamaño de espacio de búsqueda $O((b_S^n)^{n/2})$ para la p(s)-biblioteca. Lo anterior da como resultado un espacio de búsqueda de tamaño $O(b_S^n)$, que es el tamaño original del espacio de búsqueda, es decir, el ahorro se desvanece. Sin embargo, ¿qué ocurre si el número real de rotámeros de pares en los conjuntos de datos reales necesarios para lograr una cobertura de los ángulos de torsión de, por ejemplo, 90 % o más en χ_1 hasta χ_4 no es $O(b_S^2)$ sino $O(b_S)$? Si este es el caso, entonces lograremos una reducción exponencial del tamaño del espacio de búsqueda. Para lograr este objetivo consideremos que el tamaño del espacio de búsqueda $O(b_S)$ es el número de rotámeros de pares de residuos, los ángulos de torsión de residuos consecutivos necesitan estar agrupados en torno a valores específicos. A continuación se propone un algoritmo que trabaja basado en el supuesto de existencia de agrupación.

V.4. Biblioteca de pares: un enfoque basado en agrupamientos (p-biblioteca)

El enfoque propuesto para crear la p-biblioteca consta de dos fases:

1. **Organización del conjunto de datos (CD):** como se indicó anteriormente (sección V.1), se seleccionan las proteínas y de esta selección se agrupan los ángulos de torsión de las cadenas laterales por pares de residuos, es decir, el filtrado es realizado por pares de residuos. Por ejemplo, en el conjunto de datos CD149 se identificaron 17 pares de Arg-Cys, luego se toman los cuatro ángulos de torsión de

Arg (R) y se concatenan con el ángulo de torsión único de la Cys (C), obteniendo una matriz de 17 x 5, que conforma el grupo de ángulos de torsión para el par Arg-Cys. Esto se hace para cada par desde Ala-Arg hasta Val-Val.

2. **Generar un conjunto de rotámeros de pares:** esta fase consiste en la generación de los rotámeros de pares basados en el conjunto de pares de ángulos de torsión de las cadenas laterales de los residuos contiguos en el conjunto de datos (utilizando k -modas se itera hasta un umbral predefinido de la cobertura de los ángulos de torsión). Posteriormente, los rotámeros se ordenan de acuerdo con el residuo inicial en la resultante p -biblioteca.

V.4.1. Algoritmo principal

La figura 14 muestra el diagrama de flujo del algoritmo para generar los rotámeros de pares propuesto. El algoritmo comienza con el primer par de residuos (RR), es decir, los ángulos de torsión de las cadenas laterales de todos los pares contiguos de Ala-Arg en el conjunto de datos. Este conjunto de pares de ángulos de torsión (PAT) ($\chi_1, \chi_2, \chi_3, \chi_4$) de Arg (ya que Ala no tienen ángulos de torsión), se agrupa utilizando el algoritmo k -modas (Huang, 1997) implementado en Matlab. Los componentes de los vectores a ser agrupados por el algoritmo de k -modas son los valores de los ángulos de torsión correspondientes a un par particular de residuos contiguos. Cada vector es asignado a su centroide más cercano, obteniendo el nuevo rotámero para este par de residuos. Es decir, los k centroides producidos por el algoritmo de k -modas serán los k nuevos rotámeros de pares (NRP) para este par específico de residuos. En la figura 15 se ilustra la distribución de los valores de los ángulos de torsión y los rotámeros de pares utilizados en el par Val-Thr. Observe que hay grupos bien definidos alrededor de ciertas combinaciones de ángulos. Esta es una motivación para la aplicación de la

metodología que aquí proponemos.

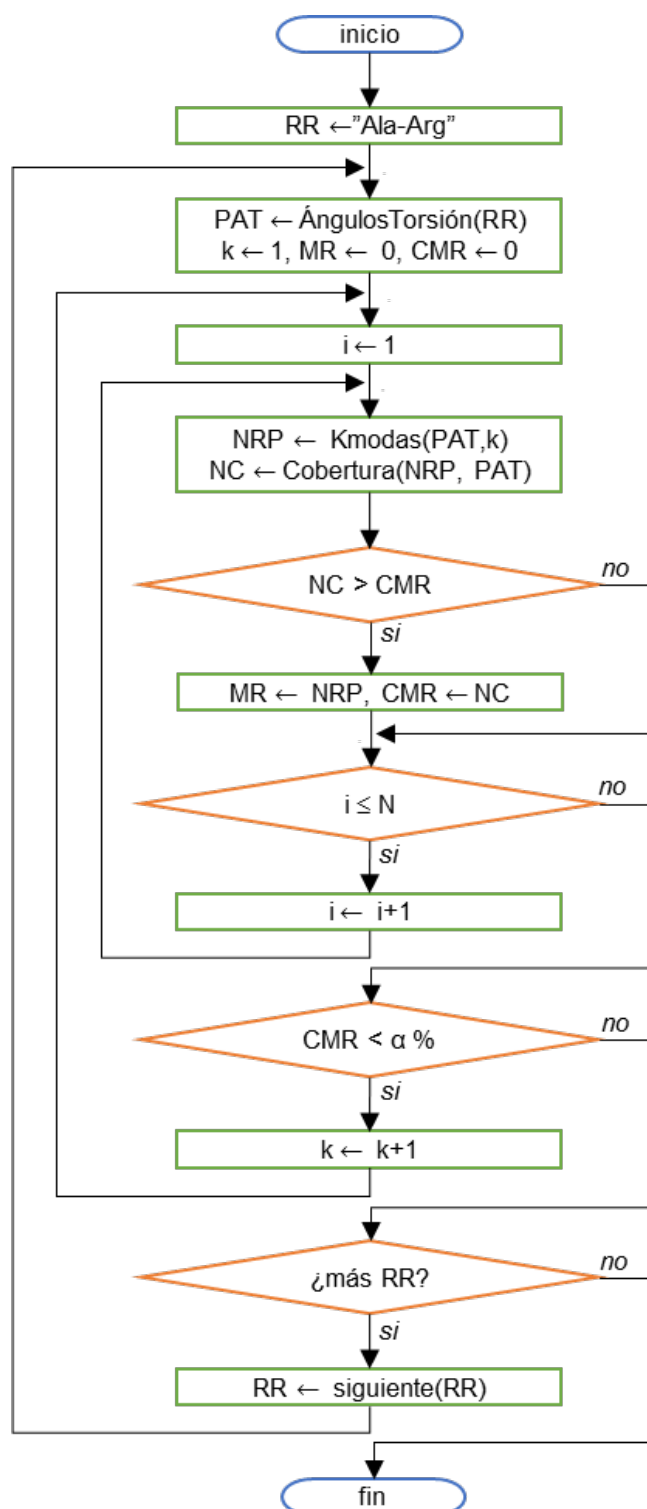


Figura 14. Algoritmo para generar los rotámeros de pares.

En el siguiente paso se calcula la nueva cobertura (NC) de los NRP generados sobre

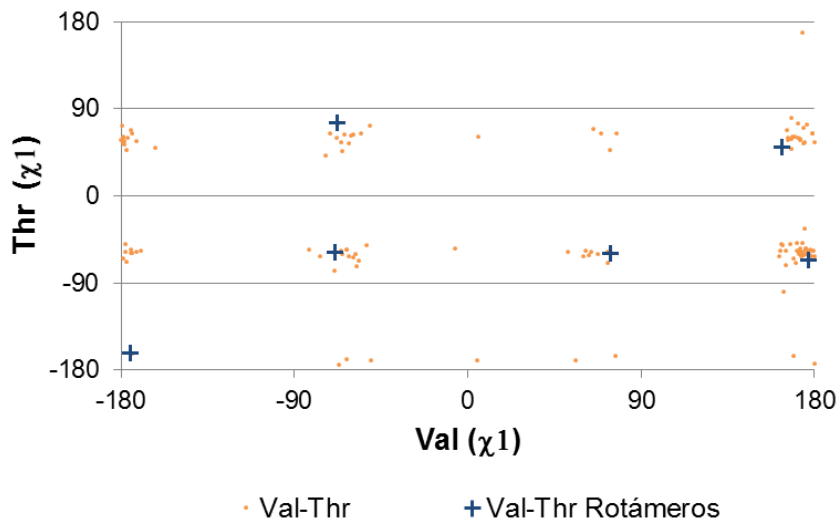


Figura 15. Ejemplo de la distribución de los ángulos de torsión y de los rotámeros en p-biblioteca.

el PAT. Para evaluar la cobertura, se calculan las diferencias entre los ángulos de torsión de la cadena lateral de cada residuo de RR (inicialmente Ala-Arg) con los ángulos de torsión de cada rotámero de pares de la biblioteca. Si al menos un rotámero cumple con la condición de que todas las diferencias sean menores a 40° , los ángulos de torsión de ese par de residuos se consideran como cubiertos por el rotámero. Para ejemplificar lo anterior, considere el caso del par específico Ala199-Arg200 de la estructura 1WTJ con ángulos de torsión $\chi_1 = -179.22^\circ$, $\chi_2 = 60.88^\circ$, $\chi_3 = 64.27^\circ$ y $\chi_4 = 66.74^\circ$, y el rotámero *R38* (para el par Ala-Arg) el cual está definido por $\chi_1 = 175.0^\circ$, $\chi_2 = 57.0^\circ$, $\chi_3 = 51.0^\circ$ y $\chi_4 = 80.0^\circ$. Vemos que las diferencias entre los ángulos de torsión en 1WTJ y los ángulos de *R38*, son: $\chi_1 = 5.78^\circ$, $\chi_2 = 3.88^\circ$, $\chi_3 = 13.27^\circ$ y $\chi_4 = 13.26^\circ$, es decir, todas las diferencias son menores a 40° , i.e., este conjunto de ángulos de torsión específicos para el par Ala199-Arg200 se considera cubierto por el rotámero *R38* y por lo tanto por la biblioteca de rotámeros. Si el NC es mayor (mejor) que la cobertura de los mejores rotámeros (CMR) hasta ahora, NC toma el lugar de CMR y NRP toma el lugar de los mejores rotámeros (MR). Debido a que k -modas depende del conjunto de

puntos iniciales (seleccionados al azar), su ejecución se repite N veces para cada valor de k . Al final, obtendremos los MR y la CMR de estas N -iteraciones. Mientras CMR sea menor que la cobertura requerida ($\alpha\%$), el valor de k se incrementa en uno y MR y CMR se vuelven a calcular como se explica en el paso anterior. Es importante tener en cuenta que este procedimiento se puede realizar de manera más eficiente utilizando un algoritmo de búsqueda binaria sobre el valor de k . Por último, se repite el proceso anterior, siguiendo una a una las parejas de los residuos hasta que todas ellas estén cubiertas (Ala-Arg, ..., Val-Val).

Algoritmo k -modas

Para la generación de rotámeros, se utilizó una técnica de agrupación conocida como k -modas (Huang, 1997), que consisten en tres pasos (ver figura 16): 1) Inicialización: se toma el grupo de residuos, se define el número de grupos (k) y un conjunto de centroides iniciales, seleccionados al azar; 2) Generación de agrupamientos: para cada vector generado por los ángulos de torsión, se calcula la distancia entre este y cada centroide y se asigna al más cercano; 3) Cálculo de centroides: para cada grupo generado en el paso anterior, se vuelve a calcular su centroide calculando la moda para cada ángulo de torsión en cada grupo. Este proceso converge cuando no hay nuevas asignaciones de ángulos en las agrupaciones. Si esta condición no se cumple, se debe regresar al paso 2. Es importante señalar que los centroides representan a los rotámeros que cubren todos los ángulos de torsión dentro de la agrupación, así como también que el conjunto de rotámeros resultantes depende de la selección inicial de centroides en el paso 1.

V.4.2. Cobertura de las bibliotecas

Como resultado de la aplicación del algoritmo propuesto para el conjunto de datos de 149 estructuras de proteínas (CD149) se obtuvo la p-biblioteca. En donde se obtienen un

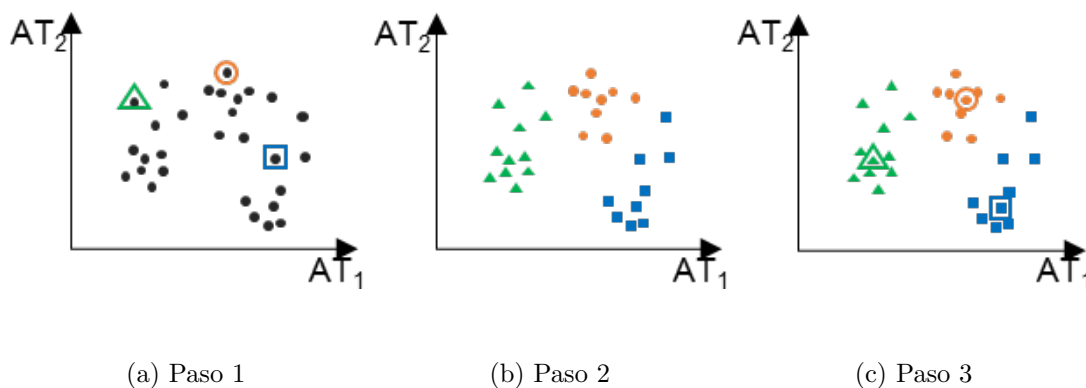


Figura 16. Algoritmo k -modas.

total de 8688 rotámeros de pares de aminoácidos consecutivos. La biblioteca generada se proporciona en la tabla del archivo p-biblioteca90.pdf. En la figura 17 se muestra el número de rotámeros, agrupados por el aminoácido inicial, generados en la p-biblioteca, este número va de 180 (sumando todos los pares que inician con Cys (C)) a 1013 (para todos los pares que inician Lys(K)) rotámeros de pares. La figura también muestra el número de pares de residuos que inician con un aminoácido dado. Las diferencias en altura de las barras para un mismo aminoácido nos dan una idea del nivel de agrupación existente entre los ángulos de torsión para todos los pares de residuos que inician con ese aminoácido. Cuanto mayor sea la diferencia entre las barras para un aminoácido dado, mayor es la agrupación que se puede lograr. Los pares de residuos que inician con Ala (A) y Gly (G) logran buenos niveles de agrupación mientras que los pares que inician con Arg (R) y Lys (K) producen agrupaciones más pobres. Un gráfico detallado para el número de rotámeros por pares de residuos se muestra en la figura 18.

Para ilustrar que el número resultante de rotámeros generados por el método propuesto no crece cuadráticamente con el número de rotámeros en una biblioteca estándar, diseñamos varias bibliotecas, estándar y de pares, para conseguir diferentes niveles de cobertura, de 70 a 100 %. La tabla VI muestra la precisión alcanzable para las bibliotecas resultantes sobre el CD149. Podemos apreciar que la biblioteca estándar puede

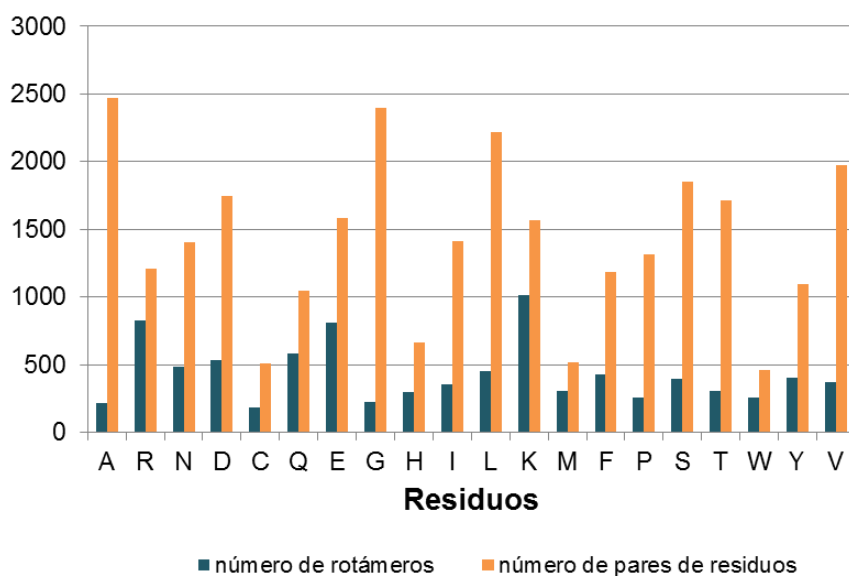


Figura 17. Rotámetros de pares de residuos en p-biblioteca y pares de residuos en CD149.

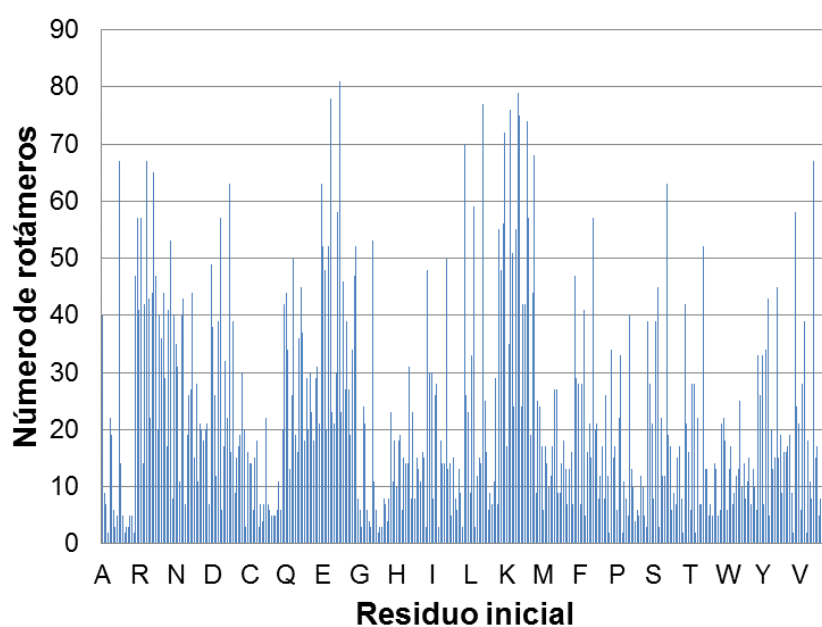


Figura 18. Número de rotámetros por pares de residuos en p-biblioteca.

lograr una mayor precisión y el costo a pagar por dicha precisión se muestra en la figura 19. La generación de una p(s)-biblioteca con un 100 % de cobertura en CD149 requiere de 4519783 rotámetros, mientras que para la p-biblioteca con el mismo porcentaje de cobertura se requiere solo 13349 rotámetros.

Tabla VI. Cobertura alcanzable para las bibliotecas resultantes sobre el CD149.

	Porcentaje de cobertura	χ_1	χ_{1+2}	χ_{1+2+3}	$\chi_{1+2+3+4}$
s-biblioteca	70 %	90 %	83 %	79 %	77 %
	80 %	96 %	90 %	87 %	86 %
	90 %	98 %	94 %	93 %	92 %
	95 %	99 %	97 %	97 %	96 %
	100 %	100 %	100 %	100 %	100 %
p-biblioteca	70 %	91 %	79 %	75 %	73 %
	80 %	95 %	87 %	84 %	82 %
	90 %	98 %	94 %	93 %	92 %
	95 %	99 %	97 %	96 %	96 %
	100 %	100 %	100 %	100 %	100 %

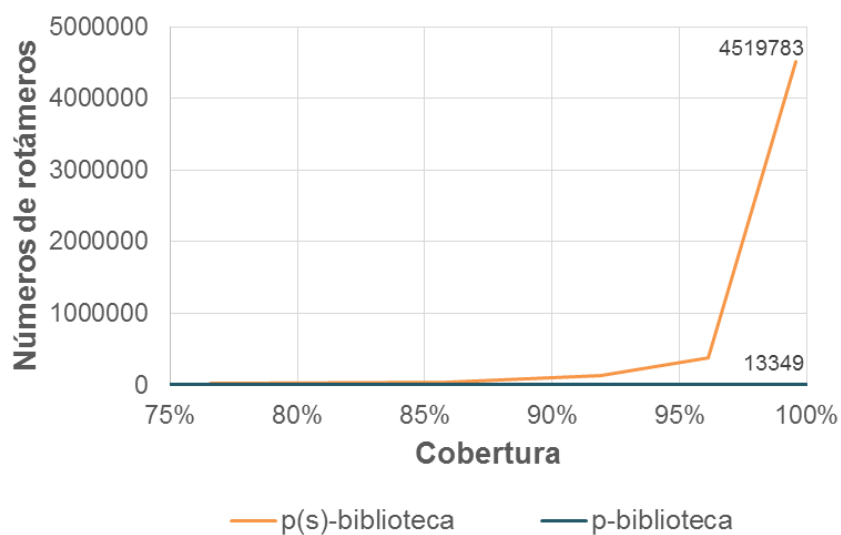


Figura 19. Número de rotámeros vs niveles de cobertura.

En la figura 20 se observa el crecimiento (en número de rotámeros) de las bibliotecas de rotámeros estándar y de pares de residuos con respecto al porcentaje de cobertura. En esta figura se evidencia que a mayor nivel de exigencia, mayor es el número de

rotámeros en las bibliotecas.

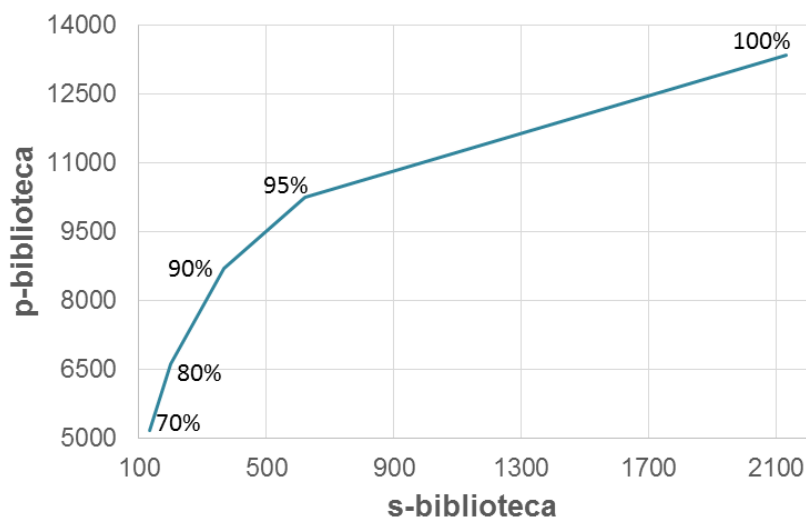


Figura 20. Comparación de tamaño de bibliotecas.

V.4.3. Correlación canónica

Después de observar la agrupación mostrada en la figura 17 por los diferentes pares de residuos, surge una pregunta natural, ¿cuál es el nivel de correlación entre los ángulos de torsión de residuos consecutivos? Para responder a esta pregunta se realizó un análisis de correlación canónica. La correlación canónica se utiliza para medir la relación entre dos conjuntos de variables aleatorias, no necesariamente de la misma cardinalidad. Es esto exactamente lo que queremos analizar, por ejemplo, si el primer residuo tiene una correlación con el segundo residuo. Para realizar este análisis se empleó la función de correlación canónica “`canoncorr`” de Matlab. En la figura 21 (a) se muestra que los residuos consecutivos que tienen un único ángulo de torsión en la cadena lateral (es decir, Cys, Ser, Thr y Val) tienen la correlación más baja, i.e., menor que 0.2. Por el contrario, en la figura 21 (c) se aprecia que los residuos consecutivos que tienen exactamente tres ángulos de torsión (Gln, Glu y Met) tienen la correlación más alta, i.e., mayor que 0.5. Los casos con la mayor correlación nos sugiere analizar la posibilidad

de implementar bibliotecas de pares de residuos, donde los ángulos de torsión para el segundo residuo se condiciona a los valores tomados por los del primer residuo. Si el número de pares de residuos con correlación es lo suficientemente grande, entonces este puede ayudar a diseñar un tipo diferente de biblioteca de pares donde los rotámetros del segundo residuo dependerán de la selección realizada por el primero. La biblioteca resultante se buscará más eficientemente que la que proponemos aquí. Por lo tanto, vale la pena hacer un análisis estadístico más profundo, el cual escapa del alcance de esta tesis.

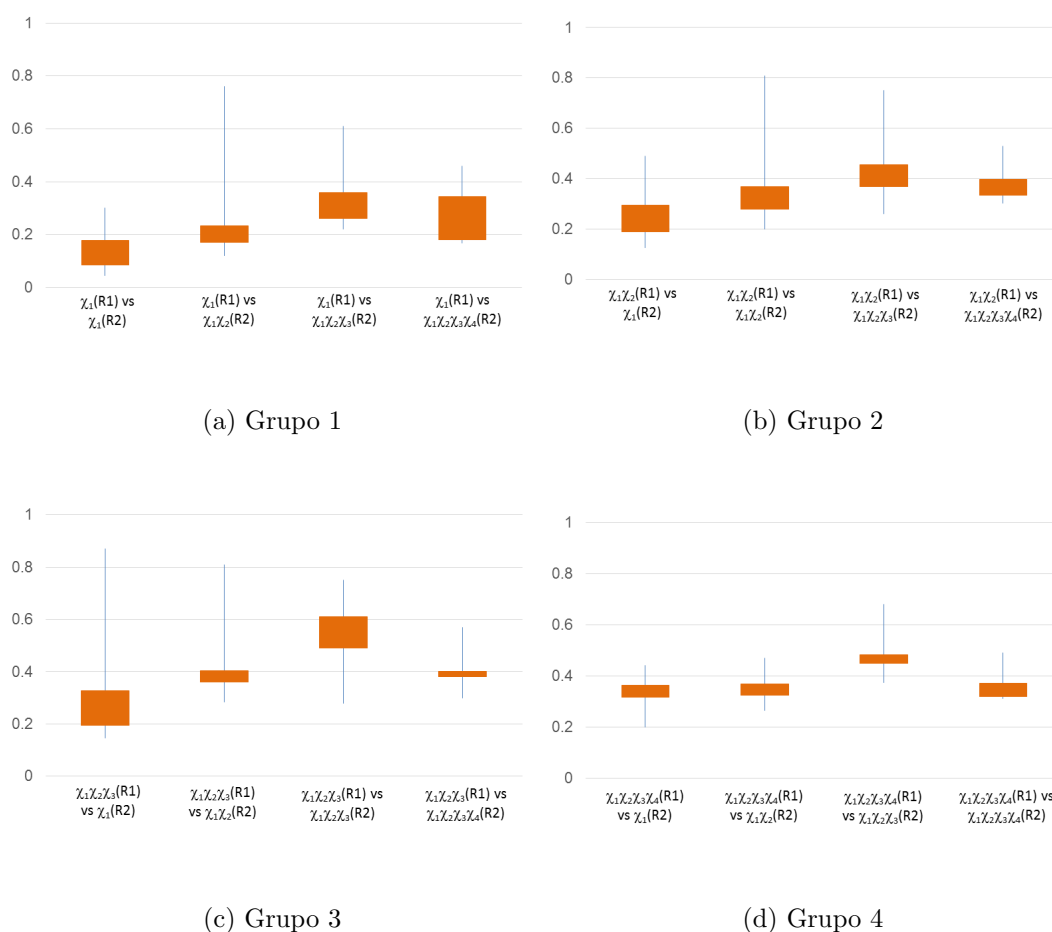


Figura 21. Correlación canónica de los ángulos de torsión del conjunto de datos CD149.

Capítulo VI

Resultados

A continuación se presentan los resultados de la evaluación de la precisión (χ_1 , χ_{1+2} , χ_{1+2+3} , y $\chi_{1+2+3+4}$) de la biblioteca de rotámeros de pares de residuos sobre dos conjuntos de prueba CP65 y CP373. Incluyendo también los resultados de la precisión reportada por las bibliotecas de Dunbrack y Lovell, así como un resumen sobre los resultados obtenidos de la evaluación de precisiones con respecto al acceso al solvente.

VI.1. Conjuntos de prueba CP65 y CP373

Para evaluar la calidad de las biblioteca utilizamos dos conjuntos de prueba. El primer conjunto está compuesto por 65 estructuras de proteínas (CP65), con las siguientes características: un único dominio bajo la clasificación SCOP, resolución de 1.2 Å o mejor, una identidad entre secuencias menor al 20 %, proteínas con secuencias de 40-631 residuos y estructuras obtenidas por cristalografía de rayos X. Este conjunto fue propuesto originalmente por Lu *et al.* (2008) para estudiar el comportamiento de un algoritmo de empaquetamiento de la cadena lateral. El segundo conjunto es de 373 estructuras de proteínas (CP373) es un subconjunto del utilizado por SCWRL4 (Krivov *et al.*, 2009). Las proteínas de este conjunto tienen una resolución de 1.8 Å o mejor, la longitud de secuencia es de 42-919 aminoácidos y el porcentaje de identidad entre las secuencias es no mayor que 30 %. Las listas de los identificadores PDB de las proteínas para cada conjunto de pruebas se proporcionan en las tablas B.1 y C.1 de los apéndices B y C, respectivamente.

En la figura 22 se muestra la abundancia de pares de residuos en los conjuntos de

prueba. En CP65, el conjunto de pares que más contribuye es el que tiene como residuo inicial Ala (A) seguida por el grupo que inicia con Gly (G); los grupos que menos contribuyen a la abundancia de pares de residuos son los que inician con Met (M) y Trp (W). En CP373, el grupo que inicia con Leu (L) es el más abundante seguido por el grupo que inicia con Ala (A); los menos abundantes son los grupos que inician con Trp (W), seguido por el grupo que inicia con Cys (C).

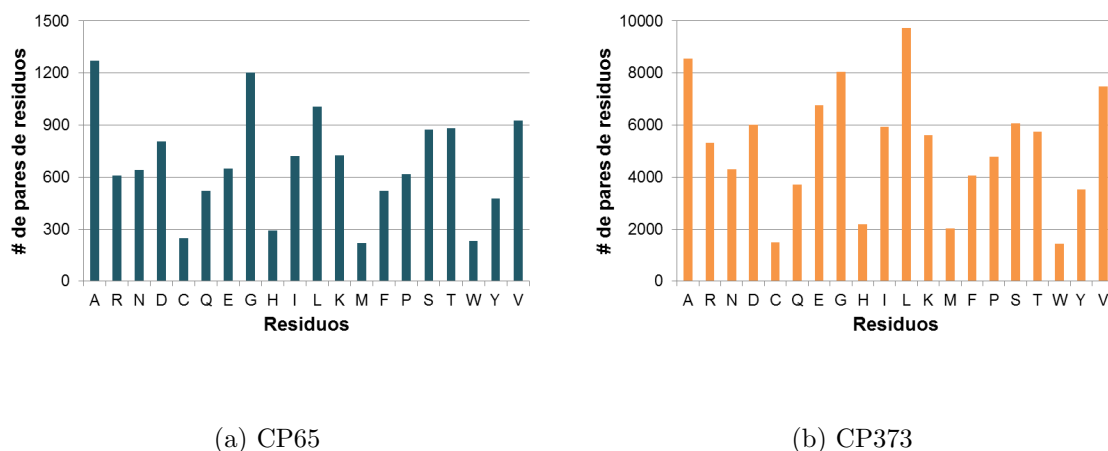


Figura 22. Abundancia de pares consecutivos de residuos en CP65 y CP373 ordenado por residuo inicial.

VI.2. Precisión alcanzable de p-biblioteca sobre CP65 y CP373

La p-biblioteca fue diseñada para lograr una cobertura de al menos 90 % en todos los ángulos de torsión de χ_1 hasta χ_4 para el conjunto de datos CD149. Sin embargo, queremos saber cómo ésta cobertura cambia cuando se utilizan diferentes tipos de estructuras. Las figuras 23 y 24 muestran el porcentaje global de precisión alcanzable de la p-biblioteca calculado sobre los conjuntos de prueba CP65 y CP373. En ambos

conjuntos se destaca el comportamiento de χ_1 en precisión alcanzable comparada con las otras tres precisiones (χ_{1+2} , χ_{1+2+3} , y $\chi_{1+2+3+4}$), en donde todos los residuos se concentran para CP65 en el intervalo de 95 % -100 % y para CP373 en el intervalo de 89 % - 100 %. En el caso de la precisión χ_{1+2} , para CP65 esta permanece por encima del 80 % (86 % en promedio) a excepción de His (H) y Trp (W) que tienen una precisión de 65 % y 70 %, respectivamente; en CP373 esta precisión (χ_{1+2}) se mantiene por encima del 66 % (83 % en promedio) a excepción de Trp (W) y His(H), ambos con 58 %. En el caso de la precisión χ_{1+2+3} , para CP65 ésta permanece por encima del 70 % (79 % en promedio) a excepción de Gln (Q), Met (M), His (H), Glu (E) y Trp (W), que tienen precisiones alcanzables de 53 %, 57 %, 61 %, 67 % y 67 %, respectivamente. Para CP373 ésta se mantiene por encima del 55 % (74 % en promedio) a excepción de Gln (Q) y Met (M), que tienen precisión alcanzable del 42 % y 41 %, respectivamente. Finalmente, en el caso de $\chi_{1+2+3+4}$, CP65 se mantiene generalmente por encima del 60 % (76 % en promedio) con las excepciones de Arg (R), Gln (Q) y Met (M), que tienen una precisión del 52 %, 53 % y 55 %, respectivamente; para CP373 estos valores se mantienen por encima del 52 % (69 % en promedio) con las excepciones de Arg (R), Gln (Q) y Met (M), que tienen una precisión del 37 %, 41 % y 39 %, respectivamente.

La figura 25 muestra el comportamiento global de los porcentajes de precisión alcanzable de la p-biblioteca calculados sobre los conjuntos de datos de diseño y los conjuntos de prueba CP65 y CP373. Los resultados del CD149 se presentan como una referencia debido a que la biblioteca fue diseñada para lograr una cobertura mínima de 90 % en CD149. Al comparar estos resultados se observa que todas las precisiones en CP65 son mejores que los de CP373 teniendo una diferencia entre los conjuntos de prueba de 0.5 %, 2.8 %, 5.4 %, y 6.5 % para χ_1 , χ_{1+2} , χ_{1+2+3} , y $\chi_{1+2+3+4}$, respectivamente.

También fueron evaluadas las p-bibliotecas generadas con cobertura del 70 %, 80 %, 90 %, 95 % y 100 % de los ángulos de torsión de χ_1 a χ_4 en los conjuntos de prueba

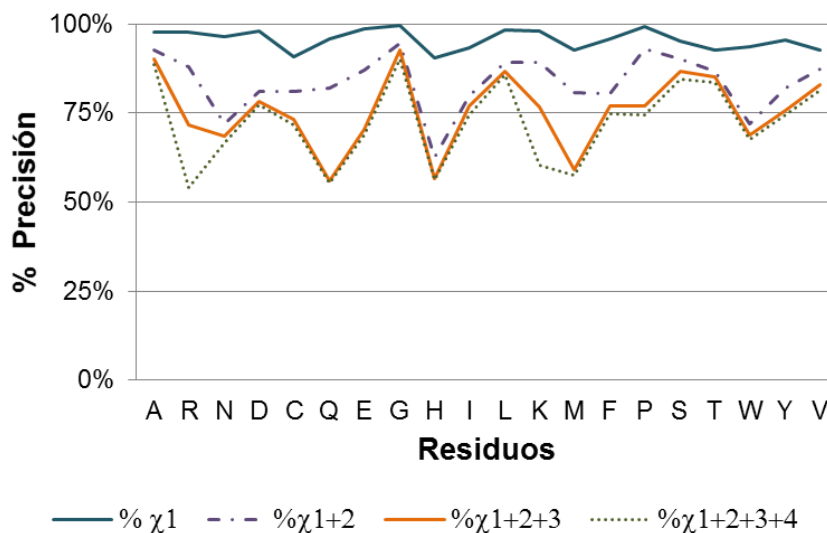


Figura 23. Porcentajes de precisión alcanzable por residuo inicial de p-biblioteca en CP65.

CP65 y CP373, cuyos resultados se encuentran en la tabla XVI del apéndice D.

Vale la pena mencionar que el número de coincidencias entre CD149 y CP65 es de 19 proteínas, mientras que entre CD149 y CP373 es de cuatro proteínas, y entre CP65 y CP373 es de seis proteínas, teniendo un porcentaje considerable de estructuras únicas, del 85 %, 97 % y 63 % para CD149, CP373 y CP65, respectivamente. Esta coincidencia entre CD149 y CP65 (21.5 % de CP65) es la razón por la cual los resultados del CP65 son mejores que los del CP373. Sin embargo, incluso con una pequeña coincidencia entre CD149 y CP373 (1 % de CP373) la precisión alcanzable de la p-biblioteca en CP373 es del 96 % para χ_1 .

En las figuras 26 y 27 se presentan los resultados del análisis de precisión por grupo, de acuerdo con el número de ángulos de torsión (AT) del CP65 y CP373. Aquí, TA- i representa todos los pares de residuos, donde el número de ángulos de torsión del residuo inicial es i . Siendo ALA (A) y GLY (G) los residuos iniciales para pares del grupo TA-0, es decir, los residuos que no tienen ángulos de torsión; Cys (C), Ser (S), Thr (T) y Val (V) son los residuos iniciales en TA-1; Asn (N), Asp (D), His (H), Ile

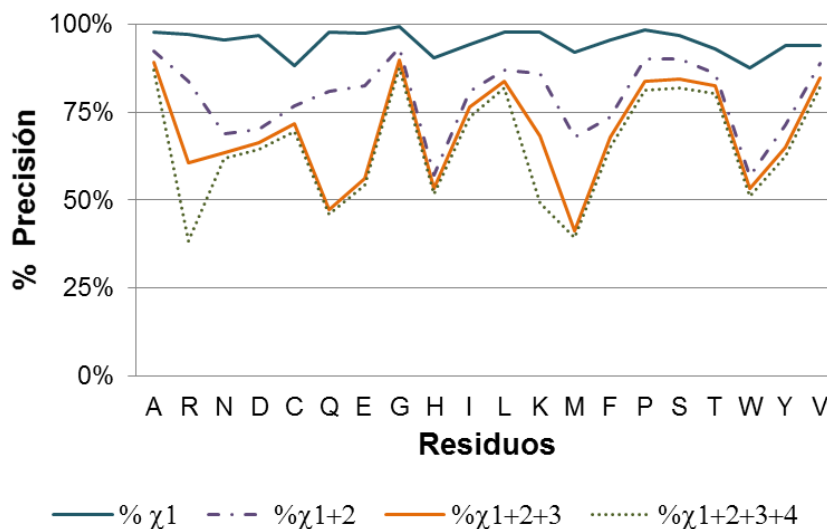


Figura 24. Porcentajes de precisión alcanzable por residuo inicial de p-biblioteca en CP373.

(I), Leu (L), Pro (P), Phe (F), Trp (W), y Tyr (Y) en TA-2; Gln (Q), Glu (E) y Met (M) en TA-3; y Arg (R) y Lus (K) en TA-4.

En CP65 el grupo TA-4 se desempeña mejor para precisión χ_1 con 98.10 %, en contraste con el grupo TA-1 obtiene solamente 93.78 % en esta precisión; para χ_{1+2} la mejor precisión fue la del grupo TA-0 con 92.80 %, en el otro extremo del grupo TA-2 que solo alcanzó el 78.96 % de precisión; para χ_{1+2+3} la mejor precisión fue la del grupo TA-0 con 90.21 %, en el otro extremo del grupo TA-3 solo alcanzó el 59.31 % de precisión; por último, para $\chi_{1+2+3+4}$, el grupo con mejor comportamiento fue TA-0 con 88.5 %, y el peor fue el grupo TA-4 con solo 56.05 % de precisión alcanzable.

En CP373 el grupo TA-0 se desempeñó mejor que los otros en todos los niveles de exigencia. Decimos que la precisión χ_{1+2} tiene mayor nivel de exigencia que la precisión χ_1 , ya que en este último caso contamos esos residuos que tienen correcto su primer ángulo de torsión. Mientras que en χ_{1+2} para contar un residuo como correcto pedimos que su primer y segundo ángulos de torsión sean correctos, el grupo TA-0 obtuvo 98.5 %, 93.6 %, 90.51 % y 88.33 % para χ_1 , χ_{1+2} , χ_{1+2+3} y $\chi_{1+2+3+4}$, respectivamente. En

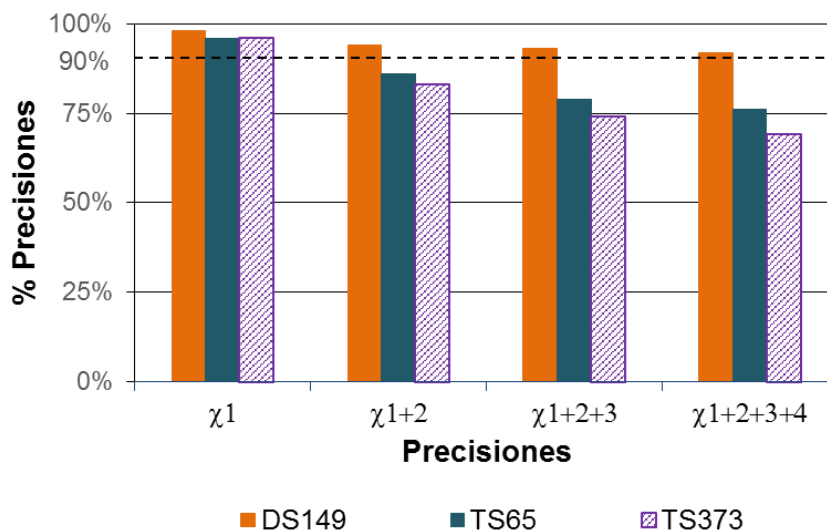


Figura 25. Porcentaje de precisión global de la p-biblioteca probado en CP149, TS65 y TS373.

contraste, el peor porcentaje obtenido para χ_1 fue por el grupo TA-1 con 93.91 %, para χ_{1+2} fue el del grupo TA-2 con 72.95 %; para χ_{1+2+3} fue el grupo TA-3 con 45.77 %, y para $\chi_{1+2+3+4}$ fue el grupo TA-4 con 42.92 %.

En general, en ambos conjuntos de prueba, el grupo TA-0 tiene una precisión mayor que los otros grupos, en todos los niveles de exigencia, es decir, hasta $\chi_{1+2+3+4}$. Esto se debe a que Ala (A) y Gly (G) no tienen cadenas laterales, entonces, solo los ángulos de torsión del segundo residuo de los pares que inician con cualquiera de estos dos residuos son considerados. Además, la diferencia en la precisión que se observa entre los niveles de exigencia dentro de los grupos se asocia al número de ángulos de torsión en el residuo inicial de la pareja. Es decir, en el caso del grupo TA-4 vemos una diferencia clara y uniforme entre niveles consecutivos de exigencia; en contraste TA-0 muestra pequeñas diferencias entre niveles consecutivos. Esto es una consecuencia del número de ángulos de torsión implicados. Es decir, a menor número de ángulos de torsión que intervienen en la cadena lateral (es decir, cadenas laterales más cortas) es más pequeña la diferencia en la precisión alcanzable entre niveles consecutivos de exigencia.

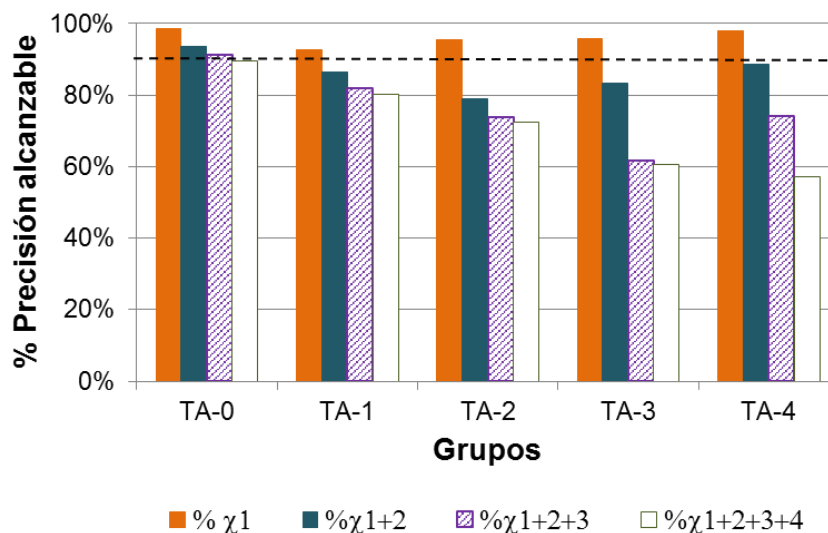


Figura 26. Porcentaje de precisión alcanzable global de la p-biblioteca en CP65 por grupo.

Los resultados generados con CP65 y CP373 son similares a aquellos obtenidos utilizando los grupos formados de acuerdo con el número de ángulos de torsión en el primer residuo de la pareja. Las diferencias observadas se atribuyen al porcentaje de similitud de los conjuntos de prueba con el CD149.

Para analizar cómo p-biblioteca se compara con otras bibliotecas, calculamos la precisión máxima alcanzable para dos bibliotecas. Una de ellas es independiente de la cadena principal y otra dependiente de la cadena principal. Los resultados se resumen en la tabla VII. En esta tabla podemos observar que la precisión alcanzable más alta es para la biblioteca de Dunbrack independiente de la cadena principal (Dunbrack *et al.*, 1997), a excepción de χ_1 , donde la biblioteca dependiente de la cadena principal (Dunbrack y Karplus, 1993) es ligeramente superior. La p-biblioteca tiene la precisión alcanzable más baja. Sin embargo, debemos considerar lo siguiente: i) a pesar de que la p-biblioteca cuenta con la precisión alcanzable más baja, ésta es aún mejor que lo que se logra con métodos del estado del arte (véase la subsección siguiente), y ii) A pesar de que la biblioteca independiente de la cadena principal muestra la mayor exactitud

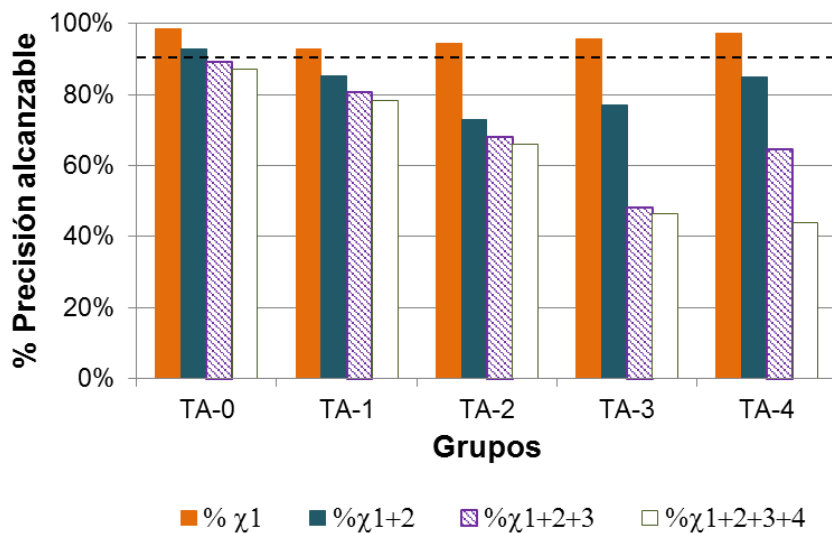


Figura 27. Porcentaje de precisión alcanzable global de la p-biblioteca en CP373 por grupo.

posible, ningún método del estado del arte la utiliza. Lo anterior se debe a que la biblioteca de rotámeros es solo uno de los componentes en el método de predicción que tiene que ser combinada con la función de energía y los algoritmos de optimización para explotar las mejores características de cada uno de ellos.

VI.2.1. Precisión reportada

Los experimentos en el CP65 utilizando el método del OPUS-Rota (Lu *et al.*, 2008) para PSCPP, mostraron una precisión de 86.7% para χ_1 , 69.7% para χ_{1+2} , 43.3% para χ_{1+2+3} y 31.7% para $\chi_{1+2+3+4}$. Por lo tanto, consideramos que con el uso de la p-biblioteca propuesta, este tipo de métodos tienen la posibilidad de aumentar sus exactitudes. Esto se debe a que la p-biblioteca puede alcanzar una precisión máxima de 96.7%, 85.5%, 78.9% y 75.8% para χ_1 , χ_{1+2} , χ_{1+2+3} y $\chi_{1+2+3+4}$, respectivamente, para el mismo conjunto de prueba. Incluso para el conjunto de prueba más grande (CP373) las precisiones alcanzables están en el nivel de 96.2%, 82.7%, 73.5% y 69.3% para χ_1 , χ_{1+2} , χ_{1+2+3} y $\chi_{1+2+3+4}$, respectivamente.

Tabla VII. Precisión alcanzable por diferentes bibliotecas.

Conjunto de prueba	Biblioteca	χ_1	χ_{1+2}	χ_{1+2+3}	$\chi_{1+2+3+4}$
CP65	Dunbrack independiente	99.5 %	94.3 %	92.8 %	91.8 %
	Dunbrack dependiente	99.6 %	93.3 %	91.9 %	87.5 %
	Lovell	99.3 %	91.4 %	88.4 %	87.5 %
	p-biblioteca	96.1 %	85.7 %	78.7 %	75.6 %
CP373	Dunbrack independiente	99.6 %	94.7 %	92.5 %	92.0 %
	Dunbrack dependiente	99.7 %	93.6 %	91.4 %	91.8 %
	Lovell	99.4 %	90.4 %	87.1 %	86.5 %
	p-biblioteca	96.1 %	82.5 %	73.4 %	69.2 %

Es importante resaltar que los resultados presentados no implican que las bibliotecas que se utilizan en los métodos del estado del arte no puedan alcanzar precisiones comparables con las aquí presentadas. En este sentido, una biblioteca ampliamente utilizada (Dunbrack dependiente de la cadena principal), véase la tabla VII, puede lograr una mejor precisión. Sin embargo, el tamaño del espacio de búsqueda será exponencialmente mayor.

Un enfoque para predecir el empaquetamiento de la cadena lateral, que ignora la función de energía y el algoritmo de optimización, consiste en seleccionar el rotámero más probable de cada aminoácido para una estructura dada. Siguiendo este enfoque se obtienen los resultados que se muestran en la tabla VIII. El empaquetamiento de las cadenas laterales realizado de esta manera, ubica a la biblioteca dependiente de la

cadena principal en primer lugar en aproximación al empaquetamiento nativo para χ_1 y χ_{1+2} . Sin embargo, la independiente de la cadena principal genera mejores resultados para χ_{1+2+3} y $\chi_{1+2+3+4}$.

Tabla VIII. Predicción de la cadena lateral mediante la selección del rotámero más probable.

Conjunto de prueba	Biblioteca	χ_1	χ_{1+2}	χ_{1+2+3}	$\chi_{1+2+3+4}$
CP65	Dunbrack independiente	55.60 %	44.80 %	40.70 %	39.30 %
	Dunbrack dependiente	73.10 %	52.70 %	25.90 %	15.80 %
	Lovell	57.70 %	44.80 %	40.70 %	39.40 %
	p-biblioteca	42.20 %	31.40 %	27.80 %	26.80 %
CP373	Dunbrack independiente	56.60 %	45.80 %	40.70 %	39.30 %
	Dunbrack dependiente	73.10 %	53.30 %	23.70 %	16.50 %
	Lovell	58.40 %	44.50 %	39.80 %	38.40 %
	p-biblioteca	41.10 %	29.50 %	25.60 %	24.50 %

También analizamos la precisión alcanzable de residuos enterrados frente a residuos expuestos (con acceso al solvente) y los resultados se presentan en la tabla IX. En esta tabla observamos que en ambas bibliotecas (s y p), como era de esperarse, los residuos enterrados obtuvieron precisiones alcanzables mejores que a los residuos obtenidos por los expuestos. Para s-bibliotecas definimos al residuo que está en la zona de transición si su superficie de accesibilidad al solvente es mayor al 20 % y menor al 50 %. Para p-biblioteca enterrado significa que la superficie de los dos residuos del par tiene accesi-

bilidad al solvente en 20 % o menos, expuesto que la superficie de ambos residuos tiene acceso al solvente en 50 % o más, y en zona de transición que la superficie de ambos residuos del par expuesta al solvente se encuentran entre 20 % y 50 %. El término otros que aparece en la tabla IX, solo se aplica a p-biblioteca y se refiere a aquellos casos en los que un residuo del par está enterrado y el otro expuesto, o cuando uno está en la zona de transición y el otro está enterrado o expuesto. Un resultado digno de mención es que la diferencia en la precisión alcanzable entre los residuos expuestos en las dos bibliotecas es sensiblemente superior a la diferencia entre los residuos enterrados.

Tabla IX. Precisión alcanzable de residuos enterrados vs. residuos expuestos.

Conjunto de prueba	Biblioteca	Enterrado	Expuesto	En zona de transición	Otros
CP65	s-biblioteca	90 %	86 %	89 %	-
	p-biblioteca	81 %	68 %	72 %	75 %
CP373	s-biblioteca	92 %	89 %	91 %	-
	p-biblioteca	79 %	56 %	61 %	69 %

Se requiere una nota final con respecto a cómo usar la biblioteca resultante. Supongamos que WKRAV es una secuencia de residuos de una proteína que necesitamos empaquetar. Usaremos rotámeros para W-K, a continuación, para la R-A, y, finalmente, vamos a considerar todos los rotámeros en V-X, donde X es cualquiera de los 18 residuos restantes (Gly y Ala no están incluidos). Es decir, si la longitud de la secuencia es impar, el último residuo tendrá que considerar todos los rotámeros de cada par en donde participa.

En resumen, la p-biblioteca generada fue probada en dos conjuntos de prueba CP65 y CP373, mostrando un promedio de precisión alcanzable para χ_1 de 96.7 % y 96.2 %, respectivamente. Se analizó también la precisión máxima de la p-biblioteca con respecto a las bibliotecas independientes de la cadena principal de Dunbrack y Lovell. Se observó

que aunque la p-biblioteca obtuvo la precisión más baja, esta es aún mejor que lo que se logra con métodos de predicción de plegado de cadenas laterales de proteínas del estado del arte. Finalmente, en el análisis de precisión bajo el criterio de exposición al solvente, encontramos que los pares de residuos enterrados fueron los que obtuvieron mejor precisión (81 % y 79 % para los conjuntos de prueba CP65 y CP373, respectivamente).

Capítulo VII

Conclusiones y trabajo futuro

En este capítulo se presenta un resumen de lo más relevante realizado en el trabajo de tesis, así como las conclusiones. Se presentan también algunas perspectivas de investigación sobre el problema de diseño de bibliotecas de rotámeros abordado.

VII.1. Sumario

El problema de diseño de bibliotecas de rotámeros independientes de la cadena principal aquí abordado, consiste en generar un multiconjunto de rotámeros de mínima cardinalidad. Este conjunto de estructuras se caracteriza por cubrir en un porcentaje específico al conjunto de residuos (*i.e.*, a sus respectivos ángulos de torsión) que le dieron origen. En este trabajo se proponen dos métodos para la generación de bibliotecas de rotámeros. El primero de ellos consiste en la generación de bibliotecas estándar (s-biblioteca), es decir, un conjunto de rotámeros por cada residuo, así como también un método para la generación de bibliotecas de pares, es decir, un conjunto de rotámeros para cada par de aminoácidos con un enfoque naïf (p(s)-biblioteca). El segundo método consiste en el diseño de bibliotecas de rotámeros que considera un enfoque de agrupamiento de ángulos de torsión. En este método primeramente se selecciona de un conjunto de estructuras de proteínas y posteriormente se generan los rotámeros de pares de residuos mediante el algoritmo conocido como k-motas.

Para analizar el crecimiento del número de rotámeros como función de la cobertura requerida, se generan p-bibliotecas con coberturas del 70 %, 80 %, 90 %, 95 % y 100 % de los ángulos de torsión de χ_1 a χ_4 , para un conjunto de datos específico de 149 estructuras de proteínas seleccionadas de la literatura. Así mismo, las bibliotecas generadas fueron probadas en dos conjuntos de prueba de 65 y 373 estructuras de proteínas, sobre las cuales se analizó el promedio de precisión alcanzable para χ_1 , χ_{1+2} , χ_{1+2+3} y $\chi_{1+2+3+4}$. Con el objeto de analizar el grado de relación de los ángulos de torsión del residuo i -ésimo con los del residuo $i + 1$, también se realiza el cálculo de la correlación canónica.

VII.2. Conclusiones

A continuación se presentan las conclusiones obtenidas de nuestro estudio acerca del problema del diseño de bibliotecas de rotámeros de nuestra investigación.

- El tamaño de las bibliotecas, tanto estándar como de pares, se incrementa a medida que incrementamos el nivel de exigencia. El crecimiento de la p-biblioteca es más lento que el de p(s)-biblioteca.
- El tamaño de la biblioteca de pares crece como una función lineal del tamaño de la biblioteca estándar. Esto es la clave para lograr una reducción exponencial en el tamaño del espacio de búsqueda y se debe al agrupamiento de los ángulos de torsión de las cadenas laterales.
- Dependiendo del número de ángulos de torsión de la cadena lateral del residuo o de las cadenas laterales del par de residuos será el número de rotámeros que aporta a la biblioteca, es decir, a mayor número de ángulos de torsión, mayor número de rotámeros en la biblioteca.

- El porcentaje de precisión alcanzable con nuestra p-biblioteca para el conjunto de prueba CP65 es mejor que el que se obtiene al utilizar OpusRota, uno de los métodos del estado del arte para la predicción del empaquetamiento de cadenas laterales. Esto nos permite conjeturar que si la p-biblioteca se incluye en estos métodos, sería posible lograr una buena precisión y con un menor tiempo de cómputo.
- Tanto la biblioteca estándar como la de pares, logran mayor precisión alcanzable sobre los residuos enterrados que sobre residuos expuestos. Se conjetura que esto se debe a que la mayoría de los residuos enterrados tienen pocas conformaciones las cuales son cubiertas en mayor porcentaje por los rotámeros. A diferencia de los residuos expuestos al solvente, que tienen mayor flexibilidad, por lo que su número de conformaciones es mayor, y el porcentaje de cobertura es menor.
- La biblioteca con mayor precisión alcanzable para los conjuntos de prueba CP65 y CP373 es la dependiente de Dunbrack. Sin embargo, es importante recordar que la biblioteca de rotámeros es solo uno de los componentes en el método de predicción. Debemos recordar que ésta debe de combinarse con la función de energía y los algoritmos de optimización para explotar las mejores características de cada uno de estos componentes.

VII.3. Propuesta de trabajo futuro

Una de las secuelas interesantes del presente trabajo es el desarrollo o adaptación de un método de PSCPP para que utilice nuestra biblioteca de rotámeros de pares de residuos. Esta propuesta está motivada por la hipótesis de que este tipo de bibliotecas pueden reducir el tiempo de cálculo en problemas de empaquetamiento de las cadenas

laterales en proteínas.

Así mismo, se propone adaptar el algoritmo propuesto en este trabajo de tesis para generar una biblioteca de rotámeros de pares de residuos pero que sea dependiente de la cadena principal. Lo anterior está motivado por los resultados obtenidos al comparar las bibliotecas (independientes y dependientes) de la cadena principal con la biblioteca de rotámeros de pares de residuos independientes de la cadena principal.

Motivados por los resultados de la correlación canónica, se planea evaluar la posibilidad de implementar una biblioteca de rotámeros de pares de residuos en donde los ángulos de torsión para el segundo residuo se condicionan a los valores tomados por los ángulos del primer residuo. Con lo anterior, se pretende lograr que la biblioteca de rotámeros de pares de residuos resultante sea explorada más eficientemente que la propuesta en este trabajo.

Finalmente, proponemos una secuela que está relacionada con el código implementado en este trabajo de tesis. Dicho código cuenta con varias funciones y programas implementados por separado como son: la generación de bibliotecas individual y de pares con la opción de elección del grado de cobertura deseado, el cálculo de precisiones a elegir por nivel de exigencia, la evaluación de bibliotecas por criterios de mejor probabilidad de rotámero o al azar, el cálculo de correlación canónica, entre otras. La propuesta es que se desarrolle una plataforma que permita a unir estos programas y que a través de esta permita a los usuarios generar sus bibliotecas de manera amigable.

Bibliografía

- Akutsu, T. (1997). Np-hardness results for protein side-chain packing. *Genome Informatics*, **8**: 180–186.
- Bhuyan, M. y Gao, X. (2011). A protein-dependent side-chain rotamer library. *BMC Bioinformatics*, **12**: S10.
- Brändén, C. y Tooze, J. (1999). *Introduction to Protein Structure*. Introduction to Protein Structure Series. Garland Pub., segundo edición. ISBN 9780815323051.
- Bryant, S. y Altschul, S. (1995). Statistics of sequence-structure threading. *Current Opinion in Structural Biology*, **5**(2): 236–244.
- Cao, Y., Song, L., Miao, Z., Hu, Y., Tian, L., y Jiang, T. (2011). Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. *Bioinformatics*, **27**(6): 785–790.
- Chan, H. y Dill, K. (1993). The protein folding problem. *Physics Today*, **37**: 24–32.
- CienciaExplicada (2015). Ciencia explicada, dogma central de la biología molecular. Recuperado el 15-10-2014 de: <http://www.ciencia-explicada.com/2009/08/video-del-dogma-central-de-la-biologia.html>.
- CNX, O. (2015). Openstax cnx, representing proteins in silico and protein forward kinematics. Recuperado el 15-02-2015 de: http://cnx.org/resources/cfe2f549f5a8e9c059e3be6b3072af76948c0144/dihedrals_illustration.png.
- Corona de la Fuente, R. (2010). *Análisis comparativo de dos Heurísticas para el Problema de Empaquetamiento de la Cadena Lateral en Proteínas*. Tesis de maestría, Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, México.
- Creighton, T. (1993). *Proteins: structures and molecular properties*. Freeman, segundo edición. ISBN 0716723174.
- CSIC (2015). Csic, department of crystallography and structural biology. Recuperado el 25-03-2015 de: http://www.xtal.iqfr.csic.es/Cristalografia/archivos_07/esquema-resolucion.jpg.
- De Maeyer, M., Desmet, J., y Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding & design*, **2**: 53–66.
- De Silva, M. (2007). Nanotecnología y nanomedicina: un nuevo horizonte para el diagnóstico y tratamiento médico. *Archivos de la Sociedad Española de Oftalmología*, **82**: 333–334.

- DeLano y Warren, L. (2002). The pymol molecular graphics system.
- Dunbrack, R., Dunbrack, R. J., y Cohen, F. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein science : a publication of the Protein Society*, **6**(8): 1661–1681.
- Dunbrack, R. J. (2002). Rotamer libraries in the 21 st century. *Current Opinion in Structural Biology*, **12**: 431–440.
- Dunbrack, R. J. y Karplus, M. (1993). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *Journal of Molecular Biology*, **230**: 543–574.
- Engh, R. y Huber, R. (1991). Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallographica Section A*, **47**: 392–400.
- Gainza, P., Roberts, K., y Donald, B. (2012). Protein design using continuous rotamers. *PLoS Computational Biology*, **8**: 1–15.
- Gu, J. y Bourne, P. (2009). *Structural Bioinformatics*. Methods of biochemical analysis. Wiley. ISBN 9780470181058.
- Hernández, S. y Brizuela, C. (2014). A k-means based approach for the design of rotamer libraries: Preliminary results. *6th International Conference on Bioinformatics and Computational Biology (BICoB 2014), Las Vegas. 2014*, páginas 15–20.
- Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. En *DMKD*, página 0. Citeseer.
- Irache, J. M. (2008). Nanomedicina: nanopartículas con aplicaciones médicas. En *Anales del Sistema Sanitario de Navarra*, Vol. 31, páginas 7–10. SciELO Espana.
- Janin, J., Wodak, S., Levitt, M., y Maigret, B. (1978). Conformations of amino acid side-chains in proteins. *Journal of Molecular Biology*, **125**: 357–386.
- Kono, H. y Doi, J. (1996). A new method for sidechain conformation prediction using a hopfield network and reproduced rotamers. *Journal of Computational Chemistry*, **1996**: 1667–1683.
- Krivov, G., Shapovalov, M., y Dunbrack, R. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. PMID: 19603484.
- Laboratory, R. (1999). 3d macromolecule analysis and kinemage home page. Recuperado el 06-01-2014 de: <http://kinemage.biochem.duke.edu/databases/top500.php>.
- Liu, Z., Jiang, L., Gao, Y., Liang, S., Chen, H., Han, Y., y Lai, L. (2003). Beyond the rotamer library: genetic algorithm combined with the disturbing mutation process for upbuilding protein side-chain. *Protein*, **50**: 49–62.
- Lovell, S., Word, J., Richardson, J., y Richardson, D. (2000). The penultimate rotamer library. *Proteins*, **40**: 389–408.

- Lu, M., Dousis, A., y Ma, J. (2008). OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Science*, **17**(9): 1576–1585.
- Martínez, D. y Carbajal, G. (2012). Hidróxidos dobles laminares: arcillas sintéticas con aplicaciones en nanotecnología. *Avances en química*, **7**: 87–99.
- McGregor, M., Islam, S., y Sternberg, M. (1987). Analysis of the relationship between sidechain conformation and secondary structure in globular proteins. *Journal of Molecular Biology*, **198**: 295–310.
- Miao, Z., Cao, Y., y Jiang, T. (2011). RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**(22): 3117–3122.
- Molecularsciences (2015). Molecularsciences, structural bioinformatics. Recuperado el 15-02-2015 de: http://molecularsciences.org/files/images/torsion_angles.gif.
- Osguthorpe, D. (2000). Ab initio protein folding. *Current Opinion in Structural Biology*, **10**(2): 146–152.
- PDB (2015). Rscb protein data bank. Recuperado el 25-02-2015 de: <http://www.rcsb.org/pdb/statistics/holdings.do>.
- Peterson, R., Dutton, L., y Wand, A. (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.
- Polaina, J. (2004). Estructura, función e ingeniería molecular de enzimas implicadas en la digestión de carbohidratos. *Mensaje Bioquímico*, **28**: 61–76.
- Ponder, J. y Richards, F. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, **193**: 775–792.
- Rodríguez, D. (2014). *Diseño de Heurísticas para el Problema de Empaquetamiento de la Cadena Lateral en Proteínas*. Tesis de maestría, Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, México.
- Rodríguez, R., Chinae, G., López, N., Pons, T., y Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, **14**: 523–528.
- Sáenz, H., Lareo, L., Poutou, R., Sosa, C., y Barrera, L. (2007). Predicción computacional de la estructura terciaria de la iduronato 2-sulfato sulfatasa humana. *Biomédica*, **27**(1): 7–20.
- Schrauber, H., Eisenhaber, F., y Argos, P. (1993). Rotamers: to be or not to be? an analysis of amino acid sidechain conformations in globular proteins. *Journal of Molecular Biology*, **230**: 592–612.

- Segovia, L. y Peimbert, M. (2010). Ingeniería de proteínas y evolución dirigida. *Mensaje Bioquímico*, páginas 135–141.
- Shapovalov, M. y Dunbrack, R. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**: 844–858.
- Tortora, G., Funke, B., y Case, C. (2007). *Introducción a la Microbiología*. Ed. Médica Panamericana.
- Tuffery, P., Etchebest, C., Hazout, S., y Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol Struct. Dyn.*, **8**: 1267–1289.
- UNC (2006a). Unc, biología virtual. Recuperado el 15-09-2014 de: <http://www.virtual.unal.edu.co/cursos/ciencias/2000024/images/biomoleculas/aminoacido.gif>.
- UNC (2006b). Unc, biología virtual. Recuperado el 15-09-2014 de: <http://www.virtual.unal.edu.co/cursos/ciencias/2000024/images/biomoleculas/peptidico.gif>.
- UniProt (2015). Uniprot. Recuperado el 25-02-2015 de: <http://www.ebi.ac.uk/uniprot/TrEMBLstats>.
- Venselaar, H., Krieger, E., y Vriend, G. (2009). Homology modeling. En *Structural Bioinformatics*. Wiley-Blackwell, New Jersey.
- Wagner, I. y Musso, H. (1983). New naturally occurring amino acids. *Angewandte Chemie International Edition in English*, **22**(11): 816–828.
- West, N. y Smith, L. (1998). Side-chains in native and random coil protein conformations. analysis of nmr coupling constants and χ_1 torsion angle preferences. *Journal of molecular biology*, **280**(5): 867–877.
- Word, J., Lovell, S., LaBean, T., Taylor, H., Zalis, M., Presley, B., Richardson, J., y Richardson, D. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of molecular biology*, **285**: 1711–1733.
- Xu, J. (2005). Rapid protein side-chain packing via tree decomposition. En *Research in computational molecular biology*, páginas 423–439. Springer.

Apéndice A

Conjunto de datos CD149

A.1. Lista de estructuras en CD149

1A2P, 1A62, 1AAC, 1AAY, 1ABA, 1ADS, 1AGJ, 1AH7, 1AHO, 1AIE, 1AJJ, 1AKO, 1AMM, 1AOP, 1AQZ, 1ARB, 1ARU, 1BAB, 1BEN, 1BFD, 1BGC, 1BHP, 1BKR, 1BPI, 1BRT, 1BTK, 1BTY, 1CEM, 1CEX, 1CKA, 1CNV, 1CTF, 1CTJ, 1CXC, 1CYO, 1DIF, 1DOS, 1ECO, 1EDG, 1EDM, 1ETN, 1EZM, 1FDR, 1FDS, 1FKJ, 1FLP, 1FUS, 1FVK, 1FXD, 1G3P, 1GAI, 1GCA, 1GDJ, 1GOF, 1GVP, 1HFC, 1HMT, 1HPM, 1IDO, 1ISU, 1IXH, 1JER, 1JHG, 1KAP, 1KUH, 1LKK, 1LUC, 1MCTI, 1MLA, 1MOF, 1MOL, 1MRJ, 1MSI, 1NIF, 1NLS, 1NOT, 1NOX, 1NWP, 1ONC, 1ORC, 1OSA, 1PDO, 1PEF, 1PEN, 1PHN, 1PLC, 1POA, 1PTF, 1RA9, 1RCF, 1RGE, 1RHS, 1RIE, 1SBP, 1SMD, 1STN, 1TCA, 1TFE, 1TGX, 1TTB, 1VCC, 1VHH, 1VIE, 1VJS, 1WAB, 1WHI, 1YGE, 1YVEI, 256B, 2ACT, 2ARC, 2AYH, 2BOP, 2CPL, 2CPP, 2CTC, 2CYP, 2DRI, 2END, 2ENG, 2ERL, 2HFT, 2HMZ, 2ILK, 2MCM, 2MHR, 2MSB, 2RN2, 2SN3, 2TRX, 3BTO, 3CHY, 3CYR, 3EBX, 3GRS, 3NUL, 3PTE, 3SDH, 3VUB, 451C, 4PGA, 4XIS, 5CYT, 5NUL, 5P21, 6CEL, 7RSA, 8ABP, 8RUC

Tabla X: Residuos eliminados en CD149.

PDB	No	AA	PDB	No	AA	PDB	No	AA	PDB	No	AA
1A2PA	59	Arg	1BTY	145	Lys	1EZM	29	Arg	1HPM	77	Arg
1A62A	21	Mse	1BTY	147	Ser	1EZM	108	Arg	1HPM	100	Arg
1A62A	29	Mse	1BTY	166	Ser	1EZM	128	Met	1HPM	137	Lys
1AAC	27	Lys	1BTY	167	Ser	1EZM	181	Lys	1HPM	188	Lys
1AAC	29	Lys	1BTY	169	Lys	1EZM	246	Lys	1HPM	248	Lys
1AAC	38	Lys	1BTY	170	Ser	1EZM	274	Arg	1HPM	250	Lys
1AAC	73	Lys	1BTY	204	Lys	1FDR	105	HIS	1HPM	264	Arg
1AAC	74	Lys	1BTY	222	Lys	1FDS	77	Glu	1IXH	216	Pro
1AAC	75	Glu	1BTY	236	Ser	1FDS	78	Arg	1JER	109	Asn
1AAC	99	Arg	1BTY	239	Lys	1FDS	105	Asp	1JHGAB	74	Glu
1AAC	101	Lys	1CEX	57	Ser	1FDS	202	Glu	1LUCB	35	Lys
1ADS	129	Asn	1CXC	10	Lys	1FDS	227	Arg	1MRJ	70	Tyr
1AHO	30	Lys	1CXC	32	Arg	1FDS	268	Asp	1MSI	2	Gln
1AHO	50	Lys	1CXC	33	Asn	1FDS	281	Arg	1NIF	8	Asp
1BENAB	14	Tyr	1CXC	35	Lys	1FDS	286	Asp	1NIF	54	Arg
1BFD	335	Gln	1CXC	74	Glu	1FDS	287	Val	1NIF	139	Glu
1BGC	9	Ser	1CXC	78	Gln	1FLP	16	Lys	1NIF	146	Val
1BGC	46	Glu	1CXC	85	Lys	1FLP	86	Glu	1NIF	166	Lys
1BGC	52	Arg	1CXC	89	Glu	1FLP	90	Lys	1NIF	189	Glu
1BGC	59	Gln	1CXC	95	Lys	1FLP	111	Lys	1NIF	195	Lys
1BGC	68	Gln	1CXC	97	Lys	1FVKA	13	Glu	1NIF	197	Lys
1BGC	123	Glu	1CXC	105	Lys	1FVKA	14	Lys	1NIF	208	Lys
1BGC	140	Thr	1CXC	106	Lys	1FVKA	52	Glu	1NIF	230	Asp
1BGC	142	Thr	1ECO	15	Lys	1FVKA	55	Lys	1NLS	9	Leu
1BKR	5	Lys	1ECO	29	Lys	1FVKA	148	Arg	1NLS	72	Ser
1BKR	20	Pro	1ECO	37	Lys	1FVKA	164	Gln	1ORC	21	Lys
1BKR	64	Gln	1ECO	40	Gln	1GAI	87	Ile	1PDO	60	Lys
1BKR	78	Lys	1ECO	44	Lys	1HMT	21	Lys	1PDO	85	Glu
1BKR	93	Lys	1ECO	50	Lys	1HMT	32	Val	1PDO	124	Glu
1BKR	108	Lys	1ECO	77	Glu	1HMT	58	Lys	1PDO	127	Lys
1BTKA	17	Lys	1ECO	88	Lys	1HMT	79	Lys	1PTF	83	Lys
1BTKA	18	Lys	1ECO	95	Asp	1HMT	80	Val	1RA9	17	Glu
1BTKA	89	Met	1ECO	109	Lys	1HMT	96	Lys	1RA9	52	Arg
1BTY	61	Ser	1EDMB	63	Lys	1HMT	107	Glu	1RCF	64	Ser
1BTY	109	Lys	1EDMB	65	Asp	1HMT	109	Ile	1SBP	309	Lys

Continúa en la siguiente página

Tabla X. Residuos eliminados en CD149– continuación.

PDB	No	AA	PDB	No	AA	PDB	No	AA	PDB	No	AA
1BTY	116	Ser	1EDMB	70	Glu	1HMT	110	Asp	1TFE	98	Met
1BTY	117	Arg	1EDMB	78	Glu	1HMT	112	Lys	1TTBA	38	Asp
1BTY	127	Ser	1EZM	28	Asp	1HMT	113	Leu	1TTBA	39	Asp
1TTBA	51	Glu	2CYP	35	Glu	2HMZA	72	Glu	3CYR	95	Lys
1TTBA	62	Glu	2CYP	38	Asn	2ILK	6	Thr	3CYR	102	Lys
1TTBA	80	Lys	2CYP	74	Lys	2ILK	8	Ser	3NUL	12	Mse
1TTBA	92	Glu	2CYP	90	Lys	2ILK	9	Glu	3NUL	47	Ile
1TTBA	104	Arg	2CYP	93	Glu	2ILK	10	Asn	3NUL	73	Mse
1TTBA	124	Asn	2CYP	97	Lys	3CYR	15	Gln	3NUL	110	Mse
2ACT	87	Glu	2CYP	120	Gln	3CYR	26	Glu	3SDHA	3	Val
2ARCB	87	Glu	2CYP	183	Lys	3CYR	27	Lys	3SDHA	4	Tyr
2ARCB	124	Glu	2CYP	210	Asp	3CYR	40	Lys	3SDHA	16	Lys
2ARCB	129	HIS	2CYP	212	Lys	3CYR	41	Glu	3SDHA	25	Ile
2ARCB	144	Glu	2CYP	219	Asn	3CYR	58	Lys	3SDHA	36	Leu
2ARCB	146	Arg	2CYP	226	Lys	3CYR	59	Lys	3SDHA	93	Val
2CPP	214	Lys	2CYP	243	Lys	3CYR	62	Lys	3SDHA	133	Asn
2CYP	2	Thr	2CYP	249	Lys	3CYR	71	Lys	3VUB	59	Glu
2CYP	4	Leu	2CYP	260	Lys	3CYR	73	Glu	4XIS	69	Glu
2CYP	12	Lys	2CYP	264	Lys	3CYR	75	Lys	4XIS	70	Glu
2CYP	17	Glu	2CYP	278	Lys	3CYR	92	Glu	4XIS	73	Lys
2CYP	21	Lys	2CYP	279	Asp	3CYR	94	Lys	4XIS	175	Asp
2CYP	32	Glu	2ERL	21	Glu						

Apéndice B

Conjunto de prueba CP65

B.1. Lista de estructuras en CP65

1A7S, 1A8Q, 1AGY, 1AKO, 1AMM, 1ARB, 1B9O, 1BD8 1BJ7, 1BYI, 1C5E, 1C9O, 1CBN, 1CC7, 1CEM, 1CEX, 1CHD, 1CKU, 1CTJ, 1CZ9, 1CZB, 1CZP, 1D4T, 1DHN, 1ECA, 1EDG, 1GCI, 1HCL, 1IC6, 1IFC, 1IGD, 1IXH, 1KOE, 1MLA, 1MML, 1NAR, 1NLS, 1NOA, 1NPK, 1PLC, 1QJ4, 1QL0, 1QLW, 1QNJ, 1QQ4, 1QTN, 1QTW, 1QU9, 1RCF, 1THV, 1THX, 1VFY, 1VJS, 1WHI, 2BAA, 2CPL, 2END, 2HVM, 2PTH, 2RN2, 3LZT, 5P21, 5PTI, 7RSA

Tabla XI. Conjunto de prueba 65.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
153L	A	185	0	185	1KOE	A	172	0	172
1A7S	A	221	0	221	1MLA	A	305	0	305
1A8Q	A	274	0	274	1MML	A	251	0	251
1AGY	A	197	1	196	1NAR	A	289	0	289
1AKO	A	268	0	268	1NLS	A	237	0	237
1AMM	A	174	0	174	1NOA	A	113	0	113
1ARB	A	263	0	263	1NPK	A	150	0	150
1B9O	A	123	0	123	1PLC	A	99	0	99
1BD8	A	156	0	156	1QJ4	A	256	1	255
1BJ7	A	150	0	150	1QL0	A	482	0	482
1BYI	A	224	0	224	1QLW	A	635	4	631
1C5E	A	285	0	285	1QNJ	A	240	0	240
1C9O	A	132	0	132	1QQ4	A	198	0	198
1CBN	A	46	0	46	1QTN	A	245	0	245
1CC7	A	72	0	72	1QTW	A	285	0	285
1CEM	A	363	0	363	1QU9	A	378	3	375
1CEX	A	197	0	197	1RCF	A	169	0	169
1CHD	A	198	0	198	1THV	A	207	0	207
1CKU	A	170	0	170	1THX	A	108	0	108
1CTJ	A	89	0	89	1VFY	A	67	0	67
1CZ9	A	139	1	138	1VJS	A	469	0	469
1CZB	A	141	0	141	1WHI	A	122	0	122
1CZP	A	196	0	196	2BAA	A	243	0	243
1D4T	A	115	0	115	2CPL	A	164	0	164
1DHN	A	121	0	121	2END	A	137	0	137
1ECA	A	136	0	136	2HVM	A	273	0	273
1EDG	A	380	0	380	2PTH	A	193	0	193
1GCI	A	269	0	269	2RN2	A	155	0	155
1HCL	A	294	0	294	3LZT	A	129	0	129
1IC6	A	279	0	279	5P21	A	166	0	166
1IFC	A	131	0	131	5PTI	A	58	0	58
1IGD	A	61	0	61	7RSA	A	124	0	124
1IXH	A	321	0	321					

Tabla XII. Residuos eliminados en CP65.

PDB	No	AA
1AGY	16	Arg
1QJ4	162	Pro
1QLW	3	Lys
1QLW	321	Lys
1QLW	350	Ser
1QLW	470	Lys
1QU9	126	Arg
1QU9	252	Arg
1QU9	378	Arg

Apéndice C

Conjunto de prueba CP373

C.1. Lista de estructuras en CP373

1AHO, 1ATZ, 1B2P, 1B8Z, 1B9W, 1BGF, 1BYI, 1C02, 1C48, 1CEI, 1DPT, 1DQ0, 1DV7, 1DYS, 1E5M, 1E6F, 1EDQ, 1ELK, 1ERZ, 1ES5, 1ES9, 1F41, 1F60, 1F94, 1FCQ, 1FO9, 1FPO, 1FVK, 1G61, 1G8A, 1G8Q, 1GMU, 1GO3, 1GPP, 1GQN, 1GS9, 1GSO, 1GVP, 1GXN, 1H03, 1H4A, 1H4Y, 1HCL, 1HZ6, 1HZ9, 1I4J, 1IGQ, 1IJQ, 1IJY, 1ILK, 1IU8, 1J23, 1J2A, 1J7G, 1JB3, 1JCD, 1JKS, 1JY2, 1K33, 1KMT, 1KMZ, 1KN3, 1KOE, 1KPT, 1KU8, 1KVA, 1KVE, 1KYF, 1KZQ, 1L3K, 1LBV, 1LM5, 1LTU, 1M5T, 1M6J, 1MD6, 1MF7, 1MG7, 1MIX, 1MKK, 1MML, 1MOL, 1N1J, 1N93, 1NAR, 1NM8, 1NWA, 1NXM, 1OAI, 1OAQ, 1OGM, 1OK7, 1P1X, 1P6Z, 1P9H, 1PCF, 1PDO, 1PE9, 1PGV, 1PM4, 1PXZ, 1QAH, 1QKD, 1R12, 1R29, 1R77, 1R8N, 1RFY, 1RGX, 1RL0, 1RWZ, 1RYL, 1RZ2, 1S7I, 1S7K, 1SAU, 1SH8, 1SMX, 1SNT, 1SQE, 1SRV, 1SUU, 1SWH, 1T1J, 1TJE, 1TKS, 1TP6, 1TUA, 1TUO, 1TY0, 1TZV, 1U07, 1U2H, 1U5X, 1UEB, 1UEK, 1UJN, 1UKF, 1ULN, 1ULR, 1UNP, 1USM, 1UXZ, 1UZ3, 1V05, 1V0S, 1V6T, 1V7Q, 1V8E, 1V8H, 1V8I, 1VDK, 1VE2, 1VGT, 1VH5, 1VJS, 1VPI, 1W5R, 1W7B, 1WBA, 1WD7, 1WKA, 1WKO, 1WLG, 1WLZ, 1WM3, 1WMH, 1WQJ, 1WTJ, 1WU9, 1WVH, 1WYC, 1WZ3, 1X1E, 1X2I, 1X6I, 1XDZ, 1XFK, 1XS0, 1XXO, 1Y2T, 1Y7Y, 1YAC, 1YHH, 1YN3, 1YO3, 1YPF, 1YT4, 1YTL, 1YU5, 1YW5, 1YXY, 1YZM, 1Z0C, 1Z0P, 1ZKR, 1ZO2, 1ZRS, 1ZUH, 1ZV1, 1ZVA, 1ZVT, 1ZXT, 2A35, 2A6W, 2A8F, 2AHF, 2AHN, 2B0A, 2B0J, 2B2F, 2BAY, 2BK8, 2BPD, 2BVP, 2CG7, 2CGH, 2CHC, 2CI3, 2CIU, 2COV, 2CWC, 2CWK, 2CWL, 2CWR, 2CYG, 2D4P, 2D68, 2D8E,

2DQW, 2DYU, 2E01, 2E10, 2E3Z, 2E64, 2E7A, 2E8F, 2E8G, 2E9Y, 2EBB, 2EBE, 2ECR, 2EGJ, 2EHG, 2EPI, 2ETX, 2EX0, 2F23, 2F5G, 2F6L, 2FBN, 2FBQ, 2FD5, 2FHZ, 2FJZ, 2FL4, 2FLU, 2FRG, 2FVH, 2FW7, 2G2U, 2G30, 2G40, 2G69, 2G7I, 2G7O, 2GAS, 2GDQ, 2GEC, 2GGV, 2GIY, 2GKG, 2GKV, 2GOM, 2GQV, 2GXF, 2H14, 2H2R, 2H2Z, 2H7W, 2H7Z, 2H8E, 2H8O, 2HC8, 2HLR, 2HLY, 2HOQ, 2HPL, 2HWW, 2HY5, 2HZF, 2I3F, 2I49, 2I5D, 2I6V, 2IBL, 2IC6, 2IC7, 2IIA, 2IJK, 2IP2, 2IPR, 2IRU, 2IUM, 2IXM, 2IY9, 2IZ6, 2J5Y, 2J6B, 2J71, 2J8B, 2J9W, 2JCP, 2NML, 2NNU, 2NPT, 2NRR, 2NV0, 2O0Q, 2O2K, 2O37, 2O6S, 2O6X, 2OEB, 2OHW, 2OIX, 2OL7, 2OSA, 2OTU, 2P1G, 2P38, 2P4H, 2P52, 2P5D, 2P5K, 2P65, 2P84, 2PBP, 2PBQ, 2PEF, 2PET, 2PGE, 2PKF, 2PMR, 2PND, 2PST, 2PTV, 2PV2, 2Q8O, 2QIY, 2QPW, 2QR3, 2QT4, 2R6U, 2R77, 2R99, 2RCZ, 2RFA, 2RIK, 2RJD, 2RK5, 2VC8, 2YXF, 2YYV, 2YZ1, 2Z14, 2Z1E, 2Z37, 2ZFY, 3BB7, 3BN6, 3C4S, 6XIA

Tabla XIII: Conjunto de prueba 373.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
1AHO	A	64	0	64	1YT4	A	263	0	263
1ATZ	A,B	373	0	373	1YTL	A,B,C,	631	3	628
1B2P	A,B	238	1	237		D			
1B8Z	A,B	133	13	120	1YU5	X	67	0	67
1B9W	A	87	3	84	1YW5	A	177	0	177
1BGF	A	124	0	124	1YXY	A,B	460	0	460
1BYI	A	224	0	224	1YZM	A	46	4	42
1C02	A,B	332	1	331	1Z0C	A	200	0	200
1C48	A,B,C,	345	0	345	1Z0P	A	73	0	73
	D,E				1ZKR	A,B	291	0	291
1CEI	A	85	0	85	1ZO2	A,B	248	0	248
1DPT	A,B,C	351	0	351	1ZRS	A,B	588	0	588
1DQ0	A	237	0	237	1ZUH	A	151	0	151
1DV7	A	212	1	211	1ZV1	A,B	118	10	108
1DYS	A,B	690	2	688	1ZVA	A	75	0	75
1E5M	A	411	2	409	1ZVT	A,B	491	0	491
1E6F	A,B	260	10	250	1ZXT	A,B,C,	276	0	276
1EDQ	A	540	0	540		D			
1ELK	A,B	306	0	306	2A35	A,B	416	4	412
1ERZ	A,B	606	0	606	2A6W	A,B	435	0	435
1ES5	A	260	0	260	2A8F	A,B	196	0	196
1ES9	A	206	15	191	2AHF	A,B	754	0	754
1F41	A,B	231	0	231	2AHN	A	221	1	220
1F60	A,B	530	0	530	2B0A	A	186	0	186
1F94	A	63	1	62	2B0J	A	344	0	344
1FCQ	A	306	2	304	2B2F	A	391	0	391
1FO9	A	331	0	331	2BAY	A,B,C,	340	0	340
1FPO	A,B,C	499	17	482		D,E,F			
1FVK	A,B	376	12	364	2BK8	A	97	0	97
1G61	A,B	450	0	450	2BPD	A,B	256	0	256
1G8A	A	227	0	227	2BVP	A,B,C	382	0	382
1G8Q	A,B	173	2	171	2CG7	A	90	0	90
1GMU	A,B,C	554	1	553	2CGH	A,B	480	9	471
	D				2CHC	A,B,C	489	32	457

Continúa en la siguiente página

Tabla XIII. Conjunto de prueba 373– continuación.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
1GO3	E,F,M,	564	41	523	2CI3	A	274	0	274
	N				2CIU	A	123	3	120
1GPP	A	217	12	205	2COV	D,E,F,	533	17	516
1GQN	A	252	0	252		G,H,I			
1GS9	A	144	0	144	2CWC	A	284	0	284
1GSO	A	419	0	419	2CWK	A,B	305	0	305
1GVP	A	87	0	87	2CWL	A,B	597	0	597
1GXN	A	323	0	323	2CWR	A	97	0	97
1H03	P,Q	250	0	250	2CYG	A	312	0	312
1H4A	X	173	1	172	2D4P	A	130	0	130
1H4Y	A,B	230	8	222	2D68	A,B	152	2	150
1HCL	A	294	0	294	2D8E	A	88	0	88
1HZ6	A,B,C	193	1	192	2DQW	A,B	497	2	495
1HZ9	A,B	132	1	131	2DYU	A,B	644	0	644
1I4J	A,B	220	0	220	2E01	A	444	7	437
1IGQ	A,B,C	227	9	218	2E10	A,B	439	0	439
	D				2E3Z	A,B	919	0	919
1IJQ	A,B	613	0	613	2E64	A,B	461	0	461
1IJY	A,B	244	0	244	2E7A	A,B,C	435	0	435
1ILK	A	151	3	148	2E8F	A	132	0	132
1IU8	A,B	412	0	412	2E8G	A,B	480	0	480
1J23	A	131	0	131	2E9Y	A,B	625	0	625
1J2A	A	166	0	166	2EBB	A	96	0	96
1J7G	A	144	0	144	2EBE	A,B	212	0	212
1JB3	A	127	0	127	2ECR	A,B	298	0	298
1JCD	A,B,C	152	1	151	2EGJ	A,B	252	2	250
1JKS	A	280	2	278	2EHG	A	149	0	149
1JY2	N,O,P,	278	0	278	2EPI	A,B,C,	388	0	388
	Q,R,S					D			
1K33	A	62	5	57	2ETX	A,B	389	0	389
1KMT	A,B	276	2	274	2EX0	A,B	782	12	770
1KMZ	A	125	1	124	2F23	A,B	308	0	308
1KN3	A	180	3	177	2F5G	A,B	260	0	260
1KOE	A	172	0	172	2F6L	A,B	330	22	308
1KPT	A,B	210	3	207	2FBN	A,B	304	0	304

Continúa en la siguiente página

Tabla XIII. Conjunto de prueba 373– continuación.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
1KU8	A,B,C, D,E,F, G,H	596	0	596	2FBQ	A	213	0	213
					2FD5	A	180	0	180
					2FHZ	A,B	199	0	199
1KVA	A	155	0	155	2FJZ	A	59	0	59
1KVE	A,B,C, D	280	0	280	2FL4	A	147	0	147
					2FLU	X,P	301	6	295
1KYF	A,P	252	0	252	2FRG	P	106	0	106
1KZQ	A,B	506	0	506	2FVH	A,B,C	297	8	289
1L3K	A	163	3	160	2FW7	A,B	320	0	320
1LBV	A,B	504	0	504	2G2U	A,B	430	0	430
1LM5	A,B	382	0	382	2G30	A,P,S	252	6	246
1LTU	A	284	0	284	2G40	A	164	5	159
1M5T	A	123	2	121	2G69	A	99	0	99
1M6J	A,B	520	4	516	2G7I	A	124	6	118
1MD6	A	154	1	153	2G7O	A	68	0	68
1MF7	A	194	9	185	2GAS	A,B	614	0	614
1MG7	A,B	703	1	702	2GDQ	A,B	746	0	746
1MIX	A	206	1	205	2GEC	A,B	260	3	257
1MKK	A,B	187	1	186	2GGV	A,B	213	0	213
1MML	A	251	0	251	2GIY	A,B	357	2	355
1MOL	A,B	188	0	188	2GKG	A	122	0	122
1N1J	A,B	165	10	155	2GKV	E,A,B	280	0	280
1N93	X	335	4	331	2GOM	A,B	121	4	117
1NAR	A	289	0	289	2GQV	A	59	0	59
1NM8	A	591	0	591	2GXX	A	140	0	140
1NWA	A	168	0	168	2H14	A	303	4	299
1NXM	A,B	385	0	385	2H2R	A,B	269	0	269
1OAI	A,B	68	0	68	2H2Z	A	306	0	306
1OAQ	H,L	230	0	230	2H7W	A,B	216	7	209
1OGM	X	572	0	572	2H7Z	A,B	152	0	152
1OK7	A,B,C	739	0	739	2H8E	A	120	1	119
1P1X	A,B	501	1	500	2H8O	A	283	1	282
1P6Z	R,N	200	0	200	2HC8	A	113	2	111
1P9H	A	179	3	176	2HLR	A	67	0	67
1PCF	A,B,C,	528	0	528	2HLY	A	205	2	203

Continúa en la siguiente página

Tabla XIII. Conjunto de prueba 373– continuación.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
	D,E,F, G,H				2HOQ	A	237	34	203
					2HPL	A,B	105	0	105
1PDO	A	129	0	129	2HWW	A,B,C	457	37	420
1PE9	A,B	722	2	720	2HY5	A,B,C	363	0	363
1PGV	A	167	0	167	2HZF	A,B	219	4	215
1PM4	A,B,C	349	15	334	2I3F	A,B	412	0	412
1PXZ	A,B	692	0	692	2I49	A	398	0	398
1QAH	A,B	264	0	264	2I5D	A	195	1	194
1QKD	A,B	124	0	124	2I6V	A	87	0	87
1R12	A,B	502	0	502	2IBL	A	108	0	108
1R29	A	122	0	122	2IC6	A,B	147	6	141
1R77	A,B	202	0	202	2IC7	A,B,C	555	3	552
1R8N	A	185	0	185	2IIA	A	89	0	89
1RFY	A,B	177	7	170	2IJK	A,B	115	4	111
1RGX	A,B,C	272	10	262	2IP2	A,B	660	34	626
1RL0	A	255	0	255	2IPR	A,B	251	0	251
1RWZ	A	244	0	244	2IRU	A,B	568	0	568
1RYL	A,B	306	0	306	2IUM	A,B,C	633	0	633
1RZ2	A	214	0	214	2IXM	A	300	0	300
1S7I	A	124	0	124	2IY9	A	309	2	307
1S7K	A	158	0	158	2IZ6	A,B	315	2	313
1SAU	A	114	0	114	2J5Y	A,B	122	5	117
1SH8	A,B	302	0	302	2J6B	A	109	0	109
1SMX	A,B	168	6	162	2J71	A	102	0	102
1SNT	A	352	0	352	2J8B	A	78	0	78
1SQE	A,B	207	0	207	2J9W	A,B	200	0	200
1SRV	A	145	0	145	2JCP	A	150	0	150
1SUU	A	293	5	288	2NML	A	100	4	96
1SWH	A,B,C, D	459	0	459	2NNU	A,B	221	0	221
					2NPT	A,B,C, D	375	25	350
1T1J	A,B	244	2	242					
1TJE	A	224	0	224	2NRR	A	139	1	138
1TKS	A,B	391	2	389	2NV0	A,B	378	6	372
1TP6	A	126	0	126	2O0Q	A	114	1	113
1TUA	A	189	0	189	2O2K	A,B	664	23	641

Continúa en la siguiente página

Tabla XIII. Conjunto de prueba 373– continuación.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
1TUO	A	437	0	437	2O37	A	81	0	81
1TY0	A,B,C	626	15	611	2O6S	A,B	416	0	416
1TZV	A	141	2	139	2O6X	A	306	1	305
1U07	A,B	178	1	177	2OEB	A	152	0	152
1U2H	A	96	3	93	2OHW	A,B	256	0	256
1U5X	A	130	10	120	2OIX	A	178	0	17
1UEB	A,B	362	0	362	2OL7	A,B	496	0	496
1UEK	A	268	0	268	2OSA	A	196	0	196
1UJN	A,B	673	2	671	2OTU	A,B,C,	976	0	976
1UKF	A	188	5	183		D,E,F,			
1ULN	A	82	0	82		G,H,P,			
1ULR	A	87	4	83		Q,R,S			
1UNP	A	119	1	118	2P1G	A,B	459	0	459
1USM	A	77	0	77	2P38	A,B	310	26	284
1UXZ	A,B	262	0	262	2P4H	X	310	0	310
1UZ3	A,B	199	1	198	2P52	A	196	0	196
1V05	A	96	0	96	2P5D	A	145	0	145
1V0S	A	495	13	482	2P5K	A	63	3	60
1V6T	A	249	0	249	2P65	A	185	4	181
1V7Q	A	169	0	169	2P84	A	133	0	133
1V8E	A	217	0	217	2PBP	A	255	0	255
1V8H	A,B	212	2	210	2PBQ	A,B,C	511	0	511
1V8I	A	150	0	150	2PEF	A	366	28	338
1VDK	A,B	915	0	915	2PET	A	230	0	230
1VE2	A,B	442	2	440	2PGE	A	356	0	356
1VGT	A,B	424	12	412	2PKF	A,B	665	1	664
1VH5	A,B	276	0	276	2PMR	A	76	3	73
1VJS	A	469	0	469	2PND	A	119	0	119
1VPI	A	122	0	122	2PST	X	61	1	60
1W5R	A,B	546	0	546	2PTV	A	96	0	96
1W7B	A	319	0	319	2PV2	A,B,C,	436	0	436
1WBA	A	171	6	165		D,E,F			
1WD7	A,B	509	0	509	2Q8O	A,B	250	5	245
1WKA	A	143	0	143	2QIY	A,B,C,	317	2	315
1WKO	A,B	330	7	323		D			

Continúa en la siguiente página

Tabla XIII. Conjunto de prueba 373– continuación.

PDB	Cadenas	AA			PDB	Cadenas	AA		
		Total	Eliminados	Útiles			Total	Eliminados	Útiles
1WLG	A,B	586	0	586	2QPW	A	147	3	144
1WLZ	A,B,C,	335	21	314	2QR3	A	121	0	121
	D								
1WM3	A	72	0	72	2QT4	A	95	0	95
1WMH	A,B	165	3	162	2R6U	A,B,C,	510	0	510
1WQJ	B,I	142	7	135		D			
1WTJ	A,B	669	0	669	2R77	A	190	5	185
1WU9	A,B	126	0	126	2R99	A	164	0	164
1WVH	A	126	11	115	2RCZ	A,B	160	9	151
1WYC	A	384	0	384	2RFA	A	222	0	222
1WZ3	A,B	168	9	159	2RIK	A	283	0	283
1X1E	A	239	0	239	2RJD	A	321	0	321
1X2I	A,B	136	0	136	2RK5	A	86	0	86
1X6I	A,B	176	1	175	2VC8	A	72	0	72
1XDZ	A	238	0	238	2YXF	A	99	0	99
1XFK	A	324	0	324	2YYV	A,B	447	0	447
1XS0	A,B,C	384	0	384	2YZ1	A,B	228	0	228
1XXO	A,B	286	0	286	2Z14	A	116	0	116
1Y2T	A,B	284	0	284	2Z1E	A	298	7	291
1Y7Y	A,B	135	0	135	2Z37	A,B,C,D	976	0	976
1YAC	A,B	408	0	408	2ZFY	A	226	11	215
1YHH	A	229	0	229	3BB7	A	314	0	314
1YN3	A,B	196	4	192	3BN6	A	158	0	158
1YO3	A,B,C	255	0	255	3C4S	A,B	112	1	111
1YPF	A,B	590	0	590	6XIA	A	387	7	380

Tabla XIV: Residuos eliminados en CP373.

PDB	No	PDB	No	PDB	No	PDB	No
PDB	NO	PDB	NO	PDB	NO	PDB	NO
1AHO	30	1PE9	135	2CHC	31	2IC6	52
1AHO	50	1PM4	28	2CHC	31	2IC6	67
1B2P	64	1PM4	38	2CHC	31	2IC7	1
1B8Z	4	1PM4	38	2CHC	39	2IC7	1
1B8Z	17	1PM4	40	2CHC	40	2IC7	185
1B8Z	22	1PM4	40	2CHC	40	2IJK	3
1B8Z	30	1PM4	41	2CHC	57	2IJK	6
1B8Z	41	1PM4	41	2CHC	79	2IJK	11
1B8Z	78	1PM4	42	2CHC	88	2IJK	11
1B8Z	78	1PM4	42	2CHC	102	2IP2	27
1B8Z	80	1PM4	42	2CHC	116	2IP2	39
1B8Z	80	1PM4	70	2CHC	116	2IP2	64
1B8Z	83	1PM4	82	2CHC	119	2IP2	64
1B8Z	86	1PM4	82	2CHC	119	2IP2	76
1B8Z	87	1PM4	82	2CHC	128	2IP2	82
1B8Z	87	1PM4	92	2CHC	128	2IP2	92
1B8Z	88	1PM4	118	2CHC	131	2IP2	115
1B8Z	90	1RFY	11	2CHC	131	2IP2	115
1B8Z	90	1RFY	15	2CHC	131	2IP2	137
1B9W	50	1RFY	42	2CHC	135	2IP2	139
1B9W	63	1RFY	45	2CHC	135	2IP2	141
1B9W	80	1RFY	46	2CHC	137	2IP2	144

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1C02	7	1RFY	98	2CHC	137	2IP2	187
1DQ0	72	1RFY	98	2CHC	138	2IP2	189
1DQ0	74	1RFY	99	2CHC	138	2IP2	193
1DV7	4	1RGX	9	2CHC	141	2IP2	196
1DYS	128	1RGX	9	2CHC	142	2IP2	213
1DYS	221	1RGX	10	2CHC	146	2IP2	217
1E5M	98	1RGX	13	2CHC	155	2IP2	222
1E5M	416	1RGX	14	2CHC	164	2IP2	228
1E6F	1514	1RGX	15	2CHC	164	2IP2	232
1E6F	1514	1RGX	15	2CIU	118	2IP2	242
1E6F	1544	1RGX	18	2CIU	120	2IP2	245
1E6F	1601	1RGX	18	2CIU	193	2IP2	245
1E6F	1602	1RGX	26	2COV	380	2IP2	267
1E6F	1602	1RGX	29	2COV	383	2IP2	274
1E6F	1615	1RYL	18	2COV	383	2IP2	276
1E6F	1615	1SMX	48	2COV	383	2IP2	281
1E6F	1617	1SMX	87	2COV	394	2IP2	289
1E6F	1622	1SMX	91	2COV	394	2IP2	298
1ELK	2	1SMX	91	2COV	394	2IP2	301
1ES9	5	1SMX	95	2COV	403	2IP2	303
1ES9	10	1SMX	95	2COV	403	2IP2	326
1ES9	15	1SMX	106	2COV	422	2IY9	82
1ES9	37	1SQE	92	2COV	423	2IY9	83
1ES9	59	1SUU	506	2COV	423	2IZ6	166

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1ES9	60	1SUU	527	2COV	424	2IZ6	166
1ES9	64	1SUU	543	2COV	425	2J5Y	8
1ES9	93	1SUU	578	2COV	425	2J5Y	8
1ES9	157	1SUU	648	2COV	425	2J5Y	20
1ES9	160	1SUU	659	2COV	436	2J5Y	36
1ES9	166	1T1J	118	2D68	69	2J5Y	36
1ES9	167	1T1J	118	2D68	69	2J5Y	50
1ES9	168	1TKS	204	2DQW	279	2J5Y	50
1ES9	198	1TKS	204	2DQW	279	2NML	13
1ES9	206	1TY0	1	2E01	2	2NML	41
1ES9	214	1TY0	1	2E01	58	2NML	50
1F94	20	1TY0	4	2E01	197	2NML	97
1FCQ	195	1TY0	5	2E01	245	2NPT	15
1FCQ	280	1TY0	47	2E01	334	2NPT	22
1FPO	18	1TY0	48	2E01	380	2NPT	34
1FPO	25	1TY0	123	2E01	427	2NPT	34
1FPO	26	1TY0	147	2EBB	3	2NPT	35
1FPO	28	1TY0	161	2EGJ	2	2NPT	38
1FPO	29	1TY0	167	2EGJ	2	2NPT	38
1FPO	45	1TY0	167	2EX0	51	2NPT	39
1FPO	48	1TY0	168	2EX0	83	2NPT	40
1FPO	49	1TY0	188	2EX0	83	2NPT	42
1FPO	51	1TY0	199	2EX0	89	2NPT	64
1FPO	56	1TY0	211	2EX0	126	2NPT	64

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1FPO	82	1TZV	75	2EX0	209	2NPT	69
1FPO	83	1TZV	132	2EX0	209	2NPT	71
1FPO	104	1U07	190	2EX0	250	2NPT	73
1FPO	130	1U2H	17	2EX0	313	2NPT	79
1FPO	134	1U2H	23	2EX0	326	2NPT	91
1FPO	142	1U2H	64	2EX0	326	2NPT	91
1FPO	171	1U5X	105	2EX0	327	2NPT	97
1FVK	13	1U5X	118	2F6L	38	2NPT	97
1FVK	13	1U5X	123	2F6L	41	2NPT	98
1FVK	14	1U5X	163	2F6L	54	2NPT	98
1FVK	14	1U5X	167	2F6L	62	2NPT	108
1FVK	47	1U5X	178	2F6L	62	2NPT	109
1FVK	52	1U5X	180	2F6L	67	2NPT	112
1FVK	52	1U5X	194	2F6L	74	2NPT	113
1FVK	55	1U5X	222	2F6L	75	2NPT	113
1FVK	132	1U5X	226	2F6L	79	2NPT	116
1FVK	148	1UJN	184	2F6L	87	2NPT	122
1FVK	148	1UJN	219	2F6L	90	2NRR	375
1FVK	164	1UKF	143	2F6L	95	2NV0	54
1G8Q	138	1UKF	151	2F6L	99	2NV0	94
1G8Q	280	1UKF	205	2F6L	103	2NV0	106
1GMU	38	1UKF	208	2F6L	138	2NV0	192
1GO3	5	1UKF	242	2F6L	139	2NV0	193
1GO3	12	1ULR	70	2F6L	143	2NV0	193

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1GO3	12	1ULR	73	2F6L	150	2O0Q	64
1GO3	19	1ULR	78	2F6L	153	2O2K	926
1GO3	23	1ULR	83	2F6L	168	2O2K	928
1GO3	24	1UNP	80	2F6L	173	2O2K	928
1GO3	24	1UZ3	100	2F6L	176	2O2K	937
1GO3	28	1UZ3	102	2FLU	78	2O2K	937
1GO3	28	1V0S	76	2FLU	446	2O2K	963
1GO3	29	1V0S	115	2FLU	447	2O2K	976
1GO3	40	1V0S	130	2FLU	493	2O2K	976
1GO3	40	1V0S	132	2FLU	528	2O2K	995
1GO3	52	1V0S	136	2FLU	551	2O2K	995
1GO3	54	1V0S	140	2FVH	2	2O2K	1000
1GO3	54	1V0S	147	2FVH	2	2O2K	1000
1GO3	56	1V0S	150	2FVH	2	2O2K	1007
1GO3	56	1V0S	163	2FVH	7	2O2K	1071
1GO3	57	1V0S	233	2FVH	7	2O2K	1077
1GO3	58	1V0S	243	2FVH	9	2O2K	1077
1GO3	58	1V0S	252	2FVH	9	2O2K	1112
1GO3	81	1V0S	257	2FVH	67	2O2K	1112
1GO3	89	1V0S	382	2G30	1	2O2K	1127
1GO3	91	1V0S	415	2G30	721	2O2K	1147
1GO3	91	1V8H	74	2G30	724	2O2K	1157
1GO3	94	1V8H	104	2G30	732	2O2K	1165
1GO3	94	1VE2	2	2G30	733	2O2K	1167

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1GO3	94	1VE2	231	2G30	828	2O2K	1167
1GO3	95	1VGT	27	2G40	77	2O2K	1223
1GO3	98	1VGT	64	2G40	79	2O2K	1226
1GO3	101	1VGT	67	2G40	112	2O2K	1235
1GO3	101	1VGT	83	2G40	164	2O6X	93
1GO3	104	1VGT	83	2G40	168	2OHW	116
1GO3	105	1VGT	98	2G7I	1149	2P38	5
1GO3	123	1VGT	125	2G7I	1203	2P38	5
1GO3	124	1VGT	128	2G7I	1206	2P38	6
1GO3	130	1VGT	148	2G7I	1210	2P38	39
1GO3	152	1VGT	148	2G7I	1216	2P38	39
1GO3	160	1VGT	172	2G7I	1230	2P38	55
1GO3	160	1VGT	203	2GEC	52	2P38	58
1GO3	177	1VGT	216	2GEC	63	2P38	75
1GO3	178	1VGT	218	2GEC	67	2P38	75
1GO3	178	1VGT	229	2GEC	74	2P38	76
1GO3	179	1WBA	18	2GIY	218	2P38	76
1GO3	180	1WBA	64	2GIY	220	2P38	80
1GPP	26	1WBA	70	2GIY	220	2P38	80
1GPP	39	1WBA	115	2GIY	395	2P38	123
1GPP	53	1WBA	161	2GKG	124	2P38	123
1GPP	54	1WBA	171	2GOM	105	2P38	128
1GPP	67	1WKO	7	2GOM	106	2P38	128
1GPP	97	1WKO	7	2GOM	107	2P38	148

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1GPP	113	1WKO	37	2GOM	165	2P38	148
1GPP	124	1WKO	38	2H14	32	2P38	151
1GPP	128	1WKO	64	2H14	78	2P38	151
1GPP	152	1WKO	131	2H14	245	2P38	155
1GPP	173	1WKO	132	2H14	247	2P38	155
1GPP	422	1WKO	171	2H7W	2	2P38	158
1H4A	117	1WLZ	227	2H7W	3	2P38	158
1H4Y	10	1WLZ	233	2H7W	3	2P38	159
1H4Y	20	1WLZ	233	2H7W	39	2P38	159
1H4Y	22	1WLZ	234	2H7W	44	2P52	121
1H4Y	30	1WLZ	234	2H7W	63	2P5K	5
1H4Y	39	1WLZ	238	2H7W	63	2P5K	46
1H4Y	88	1WLZ	241	2H7W	108	2P5K	56
1H4Y	88	1WLZ	241	2H8E	67	2P65	6
1H4Y	105	1WLZ	241	2H8O	268	2P65	101
1H4Y	106	1WLZ	241	2HC8	214	2P65	108
1HZ6	-2	1WLZ	252	2HC8	218	2P65	109
1HZ6	61	1WLZ	260	2HLR	40	2PEF	57
1HZ9	21	1WLZ	276	2HLY	178	2PEF	78
1IGQ	300	1WLZ	284	2HLY	191	2PEF	102
1IGQ	302	1WLZ	298	2HOQ	1	2PEF	109
1IGQ	302	1WLZ	298	2HOQ	3	2PEF	118
1IGQ	303	1WLZ	298	2HOQ	18	2PEF	135
1IGQ	303	1WLZ	298	2HOQ	21	2PEF	138

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1IGQ	303	1WLZ	305	2HOQ	41	2PEF	149
1IGQ	305	1WLZ	308	2HOQ	52	2PEF	153
1IGQ	305	1WLZ	308	2HOQ	56	2PEF	167
1IGQ	305	1WLZ	309	2HOQ	57	2PEF	172
1IGQ	306	1WLZ	309	2HOQ	89	2PEF	188
1IGQ	306	1WLZ	309	2HOQ	101	2PEF	197
1IGQ	306	1WMH	14	2HOQ	112	2PEF	201
1IGQ	315	1WMH	33	2HOQ	123	2PEF	205
1ILK	117	1WMH	54	2HOQ	132	2PEF	235
1ILK	119	1WQJ	11	2HOQ	134	2PEF	240
1ILK	160	1WQJ	16	2HOQ	135	2PEF	241
1JCD	5	1WQJ	27	2HOQ	138	2PEF	243
1JCD	19	1WQJ	37	2HOQ	149	2PEF	290
1JCD	46	1WQJ	42	2HOQ	157	2PEF	305
1JKS	109	1WQJ	50	2HOQ	158	2PEF	317
1JKS	277	1WQJ	66	2HOQ	161	2PEF	340
1K33	1	1WQJ	81	2HOQ	166	2PEF	341
1K33	3	1WVH	1660	2HOQ	168	2PEF	352
1K33	38	1WVH	1661	2HOQ	185	2PEF	374
1K33	65	1WVH	1662	2HOQ	186	2PEF	392
1K33	66	1WVH	1665	2HOQ	190	2PEF	393
1KMT	135	1WVH	1689	2HOQ	198	2PEF	423
1KMT	135	1WVH	1690	2HOQ	202	2PKF	0
1KMZ	44	1WVH	1704	2HOQ	208	2PKF	249

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1KN3	47	1WVH	1705	2HOQ	218	2PMR	5
1KN3	93	1WVH	1708	2HOQ	222	2PMR	15
1KN3	132	1WVH	1709	2HOQ	233	2PMR	78
1KN3	179	1WVH	1710	2HOQ	235	2PST	8
1KPT	17	1WVH	1724	2HOQ	236	2Q8O	49
1KPT	58	1WVH	1735	2HOQ	237	2Q8O	49
1KPT	65	1WZ3	10	2HWW	1244	2Q8O	50
1L3K	17	1WZ3	10	2HWW	1267	2Q8O	104
1L3K	44	1WZ3	27	2HWW	1284	2Q8O	147
1L3K	59	1WZ3	29	2HWW	1288	2QIY	71
1M5T	91	1WZ3	31	2HWW	1301	2QIY	71
1M5T	95	1WZ3	46	2HWW	1301	2QPW	108
1M6J	115	1WZ3	47	2HWW	1304	2QPW	109
1M6J	118	1WZ3	79	2HWW	1304	2QPW	143
1M6J	125	1WZ3	80	2HWW	1308	2R77	39
1M6J	195	1X6I	0	2HWW	1308	2R77	156
1MD6	9	1YN3	180	2HWW	1312	2R77	160
1MF7	166	1YN3	180	2HWW	1316	2R77	176
1MF7	225	1YN3	206	2HWW	1319	2R77	178
1MF7	278	1YN3	206	2HWW	1319	2RCZ	193
1MF7	281	1YTL	17	2HWW	1321	2RCZ	194
1MF7	282	1YTL	17	2HWW	1321	2RCZ	197
1MF7	293	1YTL	17	2HWW	1331	2RCZ	210
1MF7	306	1YZM	478	2HWW	1331	2RCZ	222

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1MF7	311	1YZM	480	2HWW	1336	2RCZ	224
1MF7	313	1YZM	499	2HWW	1338	2RCZ	262
1MG7	363	1YZM	501	2HWW	1343	2RCZ	262
1MIX	196	1ZRS	142	2HWW	1351	2RCZ	263
1MIX	318	1ZV1	60	2HWW	1351	2Z1E	41
1MIX	320	1ZV1	60	2HWW	1354	2Z1E	42
1MIX	322	1ZV1	61	2HWW	1365	2Z1E	64
1MKK	13	1ZV1	61	2HWW	1365	2Z1E	112
1N1J	50	1ZV1	62	2HWW	1366	2Z1E	153
1N1J	56	1ZV1	62	2HWW	1367	2Z1E	190
1N1J	57	1ZV1	63	2HWW	1367	2Z1E	289
1N1J	59	1ZV1	63	2HWW	1369	2ZFY	59
1N1J	64	1ZV1	64	2HWW	1369	2ZFY	63
1N1J	67	1ZV1	64	2HWW	1372	2ZFY	64
1N1J	114	2A35	176	2HWW	1380	2ZFY	73
1N1J	134	2A35	176	2HWW	1381	2ZFY	120
1N1J	137	2A35	196	2HWW	1381	2ZFY	133
1N1J	138	2A35	198	2HWW	1383	2ZFY	136
1N93	107	2AHN	108	2HWW	1393	2ZFY	151
1N93	109	2CGH	14	2HWW	1398	2ZFY	188
1N93	297	2CGH	14	2HWW	1402	2ZFY	240
1N93	352	2CGH	35	2HWW	1403	2ZFY	256
1P1X	1180	2CGH	89	2HZF	27	3C4S	1
1P9H	178	2CGH	123	2HZF	40	6XIA	3

Continúa en la siguiente página

Tabla XIV. Residuos eliminados en CP373– continuación.

PDB	No	PDB	No	PDB	No	PDB	No
1P9H	216	2CGH	135	2HZF	49	6XIA	7
1P9H	219	2CGH	135	2HZF	49	6XIA	9
1PDO	60	2CGH	161	2I5D	0	6XIA	22
1PDO	85	2CGH	161	2IC6	8	6XIA	72
1PDO	124	2CGH	169	2IC6	21	6XIA	84
1PDO	127	2CHC	7	2IC6	24	6XIA	131
1PE9	135	2CHC	7	2IC6	24		

Apéndice D

Misceláneos

Tabla XV. Ángulos de torsión de la cadena lateral por tipo de residuo.

Aminoácido	χ_1	χ_2	χ_3	χ_4	χ_5
Alanina (Ala, A)					
Arginina (Arg, R)	X	X	X	X	X
Asparagina (Asn, N)	X	X			
Ácido aspártico (Asp, D)	X	X			
Cisteína (Cys, C)	X				
Glutamina (Gln, Q)	X	X	X		
Ácido glutámico (Glu, E)	X	X	X		
Glicina (Gly, G)					
Histidina (His, H)	X	X			
Isoleucina (Ile, I)	X	X			
Leucina (Leu, L)	X	X			
Lisina (Lys, K)	X	X	X	X	
Metionina (Met, M)	X	X	X		
Fenilalanina (Phe, F)	X	X			
Prolina (Pro, P)	X	X			
Serina (Ser, S)	X				
Treonina (Thr, T)	X				
Tirosina (Tyr, Y)	X	X			
Triptófano (Trp, W)	X	X			
Valina (Val, V)	X				

Tabla XVI. Porcentaje de precisión global de p-bibliotecas evaluadas en CP65 y CP373.

Conjunto de prueba	Biblioteca	χ_1	χ_{1+2}	χ_{1+2+3}	$\chi_{1+2+3+4}$
CP65	p-biblioteca70	89 %	73 %	65 %	61 %
	p-biblioteca80	93 %	80 %	72 %	69 %
	p-biblioteca90	96 %	86 %	79 %	76 %
	p-biblioteca95	98 %	89 %	82 %	79 %
	p-biblioteca100	99 %	91 %	85 %	83 %
CP373	p-biblioteca70	89 %	70 %	60 %	56 %
	p-biblioteca80	93 %	77 %	67 %	63 %
	p-biblioteca90	96 %	83 %	74 %	69 %
	p-biblioteca95	97 %	86 %	77 %	72 %
	p-biblioteca100	98 %	88 %	80 %	76 %