Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Doctorado en Ciencias en Ciencias de la Computación

Empacamiento de la cadena lateral de proteínas: algoritmos, límites del desempeño y funciones de calificación

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de Doctor en Ciencias

Presenta:

José Domingo Colbes Sanabria

Ensenada, Baja California, México

2018

Tesis defendida por

José Domingo Colbes Sanabria

y aprobada por el siguiente Comité

Dr. Carlos Alberto Brizuela Rodriguez

Director de tesis

Dr. José Alberto Fernández Zepeda

Dr. Israel Marck Martínez Pérez

Dr. Andrey Chernykh

Dr. Sergio Andrés Águila Puentes

Dr. Jesús Favela Vara

Coordinador del Posgrado en Ciencias de la Computación

Dra. Rufina Hernández Martínez

Directora de Estudios de Posgrado

Resumen de la tesis que presenta José Domingo Colbes Sanabria como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la Computación.

Empacamiento de la cadena lateral de proteínas: algoritmos, límites del desempeño y funciones de calificación

Resumen aprobado por:

Dr. Carlos Alberto Brizuela Rodriguez
Director de tesis

Uno de los problemas abiertos más importantes en biología computacional consiste en predecir la estructura de una proteína a partir de su secuencia de aminoácidos. Tanto en este problema como en el inverso, el diseño de proteínas, el problema del empacamiento de la cadena lateral de proteínas (PSCPP por sus siglas en inglés) es de gran importancia. El PSCPP se modela generalmente como un problema de optimización combinatoria, donde la cadena lateral de cada residuo tiene un conjunto finito de conformaciones posibles (denominados rotámeros) obtenidos de una biblioteca. El problema consiste en seleccionar una conformación para cada residuo para minimizar una función de score dada, la cual considera las interacciones en el sistema conformado por la proteína y su entorno. Se han propuesto un gran número de métodos en las últimas dos décadas para resolver este problema NP-difícil, pero las precisiones que alcanzan se han estancado en valores considerablemente alejados de los ideales. Para determinar si pueden obtenerse mejoras, se calculó la máxima precisión alcanzable mediante una biblioteca de rotámeros simple, comparándola con las obtenidas por cinco métodos del estado del arte. Los resultados muestran una brecha significativa de mejora posible, por lo que el siguiente paso consistió en identificar las limitaciones de los métodos actuales. Trabajos previos en la predicción de estructura de proteínas y el diseño de proteínas indican que las imprecisiones de las funciones de score actuales podrían representar el principal obstáculo para alcanzar mejores resultados en estos problemas. Para demostrar que lo mismo se cumple para el PSCPP, se propuso un método de evaluación de funciones de score basado en la búsqueda local, empleándolo para evaluar el desempeño de las funciones de score de dos métodos del estado del arte. Los resultados señalan que ninguna de las dos funciones puede guiar correctamente a los algoritmos de búsqueda. Se exploró dos posibilidades para explicar este resultado negativo: (i) una incorrecta asignación de pesos a los términos de la función de score, y (ii) la influencia de la conformación restringida resultante del proceso de cristalización en las estructuras de referencia. Para analizar estas interrogantes: (i) se modeló el PSCPP como un problema de optimización biobjetivo, considerando los dos términos más importantes de las funciones de score seleccionadas; y (ii) se realizó un preprocesamiento de relajación de la estructura cristalográfica mediante dinámica molecular, para simular la proteína en el solvente y evaluar el desempeño de las mismas dos funciones de score en este nuevo entorno. Los resultados indican que: (i) independientemente de la combinación de pesos, las funciones de score actuales no

lograrán mejores resultados; y (ii) que las mismas tampoco podrán mejorar su desempeño con las estructuras relajadas. Todos estos hallazgos, sumados a los resultados de experimentos que determinaron que los algoritmos basados en la búsqueda local producen resultados competitivos para el PSCPP, refuerzan la idea de que los esfuerzos para mejorar los resultados actuales deben centrarse en el diseño de mejores funciones de score; pudiéndose utilizar para ello el método de búsqueda local propuesto en este trabajo.

Palabras clave: estructura de proteína, empacamiento de la cadena lateral, función de score, búsqueda local, dinámica molecular, biblioteca de rotámeros.

Abstract of the thesis presented by José Domingo Colbes Sanabria as a partial requirement to obtain the Doctor of Science degree in Computer Science.

Evaluation and analysis of scoring functions for the protein side-chain packing problem

Abstract approved by:

Dr. Carlos Alberto Brizuela Rodriguez
Thesis Director

One of the most challenging problems in computational biology involves predicting the structure of a protein given its amino acid sequence. The protein side-chain packing problem (PSCPP) is an important subproblem of this problem and its inverse, the protein design. The PSCPP is usually modeled as a combinatorial optimization problem, where each residue has a finite set of possible conformations (called rotamers) obtained from a rotamer library; and the problem consists of selecting a set of rotamers (one for each residue) in order to minimize a given scoring function, which considers the interactions within the system composed by the protein and its environment. During the past two decades, a large number of methods have been proposed to tackle this NP-hard problem, but their accuracies are stagnant in values considerably distant from the ideal ones. To determine if improvements could be obtained, the maximum accuracy achievable by a simple rotamer library was calculated, comparing it with those obtained by five state-of-the-art methods. The results show a significant gap for improvement, so the next step was to identify the limitations of current methods. Previous works on protein structure prediction and protein design have shown that scoring function inaccuracies may represent the main obstacle to achieving better results for these problems. To show that the same is true for the PSCPP, the quality of two scoring functions used by some state-of-the-art algorithms was evaluated. The results indicate that neither of these two scoring functions can guide the search method correctly. Two possibilities were explored to explain this negative result: (i) an incorrect weighting of the scoring functions terms, and (ii) the constrained conformation resulting from the protein crystallization process in reference structures. To analyze these questions: (i) the PSCPP was modeled as a bi-objective combinatorial optimization problem, considering the two most important terms of the selected scoring functions; and (ii) a pre-processing relaxation of the crystal structure was performed through molecular dynamics to simulate the protein in the solvent, in order to evaluate the performance of the same two scoring functions under this new environment. The results indicate that: (i) no matter what combination of weight factors are used, the current scoring functions will not lead to better performances, and (ii) they will not be able to improve performance on relaxed structures either. All these findings, added to the results of experiments that determined that local search based algorithms produce competitive results for the PSCPP, reinforce the idea that efforts to improve current results should

focus on the design of better scoring functions; and the local search method proposed in this work could be used for that purpose.

Keywords: protein structure, side-chain packing, scoring function, local search, molecular dynamics, rotamer library.

Dedicatoria

A Andrea, Gloria, José y Cristhian

Agradecimientos

Al analizar el camino recorrido en los últimos diez años, el Dr. Carlos Brizuela ha sido un pilar fundamental en mi formación académica y personal. Además de un profundo agradecimiento le tengo una gran admiración, y representa para mí un ejemplo a seguir. En momentos claves de mi vida me ha ofrecido su invaluable ayuda, sus consejos y observaciones siempre fueron acertados, y supo guiarme de la mejor manera en todo este tiempo. Ha sido un honor para mí ser su estudiante.

Mi esposa Andrea también ha tenido un rol fundamental. A pesar de haber renunciado a muchas cosas para embarcarse en este sueño de volver a CICESE para el doctorado, me acompañó con todas las ganas y optimismo pese a las dificultades que se nos presentaron en el camino. Agradezco su paciencia, comprensión, amor y apoyo incondicional.

A mi hermano Cristhian, mi mejor amigo. En mi ausencia en Paraguay tuvo que lidiar con muchísimas dificultades y situaciones muy amargas, y aún así siempre confió plenamente en mí y apoyó mis decisiones. Por lo tanto, estoy plenamente convencido de que mis logros son también suyos.

A mis padres Gloria y José, por haberme impulsado desde pequeño a dar lo mejor de mí, a confiar en mí mismo y a soñar en grande; y por brindarme todas las oportunidades que tuve hasta ahora en la vida. Con el paso del tiempo crece mi agradecimiento y admiración por todo el esfuerzo y sacrificio que realizaron por sus hijos.

Al Centro de Investigación Científica y de Educación Superior de Ensenada, y en especial al Departamento de Ciencias de la Computación, por haber hecho nuevamente mi pasantía por la institución lo más agradable posible. A los profesores, por haber contribuido en mi formación académica.

A los miembros del comité: Dr. Sergio Águila, Dr. Alberto Fernández, Dr. Israel Mar-

tínez y Dr. Andrey Chernykh; por sus valiosos aportes durante el desarrollo del trabajo de tesis.

A todos los compañeros del Departamento de Ciencias de la Computación, por todos los momentos compartidos durante mi estadía en el doctorado. Disfruté muchísimo el día a día con los compañeros del cubo de Biocomputación. También a los buenos amigos que hice en el fútbol, especialmente a los legendarios “Galácticos”, con quienes compartimos la pasión por este deporte. Una de las razones principales de haber elegido nuevamente el CICESE fue por la calidez de la gente y por lo mucho que disfruté al lado de mis compañeros en la maestría, y me alegro de no haberme equivocado en esta decisión.

A mis amigos de Paraguay, por estar siempre acompañándome aún en la distancia, especialmente en los momentos en los que más los necesitaba.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) y a la Universidad Nacional de Asunción, por brindarme el apoyo económico para realizar mis estudios de doctorado.

Al proyecto de supercómputo en Miztli LANCAD-UNAM-DGTIC-286, que me permitió realizar las simulaciones de dinámica molecular.

Tabla de contenido

Página

Tabla de contenido

Tabla de contenido

Lista de figuras

Figura Página

Lista de figuras

Lista de tablas

Lista de tablas

Capítulo 1. Introducción

La **biología computacional** es el área que consiste en el desarrollo y aplicación de métodos teóricos y de análisis de datos, modelado matemático y técnicas de simulación computacional para el estudio de sistemas biológicos, conductuales y sociales (Huerta *et al.*, 2000). A partir de esta definición puede observarse que es un área multidisciplinaria que emplea conocimientos de: matemáticas, ciencias de la computación, estadística, biología molecular, bioquímica, biofísica, entre otros. En ella, algunos problemas de interés están relacionados a la genómica (estudio de los genomas) y a la proteómica (estudio de las proteínas) (Chou, 2004); y a las redes de interacción tanto a nivel de genes como de proteínas (biología de sistemas) (Stelling *et al.*, 2001).

Posiblemente, uno de los problemas sin resolver más importantes de la biología computacional es la **predicción de estructuras de proteínas** (Moult *et al.*, 2016; Zhang, 2008). Las proteínas son macromoléculas formadas por cadenas lineales de aminoácidos y cumplen funciones cruciales en esencialmente todos los procesos biológicos (Berg *et al.*, 2012). Una proteína se pliega espontáneamente en una estructura tridimensional bien definida y estable, denominada **estructura nativa**. De acuerdo a la hipótesis de Anfinsen (1973), a esta estructura la determina la secuencia de aminoácidos que la componen. Conocer la estructura de una proteína es importante porque la función que cumple depende directamente de dicha estructura tridimensional (Berg *et al.*, 2012).

Actualmente, la determinación de la estructuras tridimensionales de proteínas de interés se logra a través de métodos experimentales, siendo los más utilizados los siguientes: (i) Cristalografía de rayos X (Berg *et al.*, 2012), (ii) Resonancia magnética nuclear (Nelson y Cox, 2004) y (iii) Crio-Microscopía Electrónica (Bai *et al.*, 2015).

El *PSI* (*Protein Structure Initiative*) (Berman *et al.*, 2008) fue un proyecto ambicioso que costó alrededor de 765 millones de dólares[1] entre los años 2000 y 2015. Esta iniciativa apuntó al desarrollo de nuevas tecnologías y métodos para reducir el costo y tiempo requeridos para la determinación de las estructuras de proteínas.

Entre los avances obtenidos[2] por el PSI se destacan:

[1] www.nigms.nih.gov/News/reports/archivedreports2009-2007/Pages/PSIAssessmentPanel2007.aspx
[2] sbkb.org

- Se determinaron aproximadamente 7000 estructuras.
- Se desarrollaron o mejoraron alrededor de 450 tecnologías.
- Se generaron aproximadamente 2300 publicaciones científicas.

A pesar de todos los avances alcanzados por el PSI y las perspectivas interesantes desde el punto de vista experimental, a continuación se expone la necesidad de resolver el problema de predicción de estructuras mediante un enfoque computacional.

1.1. Motivación de los métodos computacionales para la predicción de estructuras de proteínas

El *Protein Data Bank*[3] (PDB) es la base de datos de estructuras de proteínas más importante y actualmente almacena 134,091 estructuras[4]. La base de datos de secuencias de proteínas más importante es *UniProt*[5]. La misma se divide en *Swiss-Prot*, que contiene 555,594[6] secuencias (revisadas); y TrEMBL, con 90,050,711[7] secuencias (pendientes de revisión). Como puede observarse en la Figura 1, la diferencia entre secuencias y estructuras almacenadas es cada vez mayor.

Esta tendencia, que ha sido señalada a lo largo de los últimos años (Rodriguez *et al.*, 1998; Baker y Sali, 2001; Roy *et al.*, 2010; Moult *et al.*, 2016), se debe principalmente a las limitaciones en costo, al tiempo requerido y a las dificultades técnicas de los métodos experimentales empleados para determinar las estructuras de proteínas.

Es debido a esta situación que el problema de la predicción de estructuras de proteínas es uno de los desafíos más importantes dentro de la Biología Computacional, y desde hace varias décadas se buscan métodos *in silico* para determinar la estructura tridimensional de la proteína a partir de su secuencia de aminoácidos (Levinthal, 1968; Moult, 2005). El término *in silico* se emplea para indicar que algo se hace por medio de computadoras o a través de simulaciones computacionales (Palsson, 2000).

[3] www.rcsb.org
[4] Al 01/10/2017
[5] www.uniprot.org
[6] Al 01/10/2017
[7] Al 01/10/2017

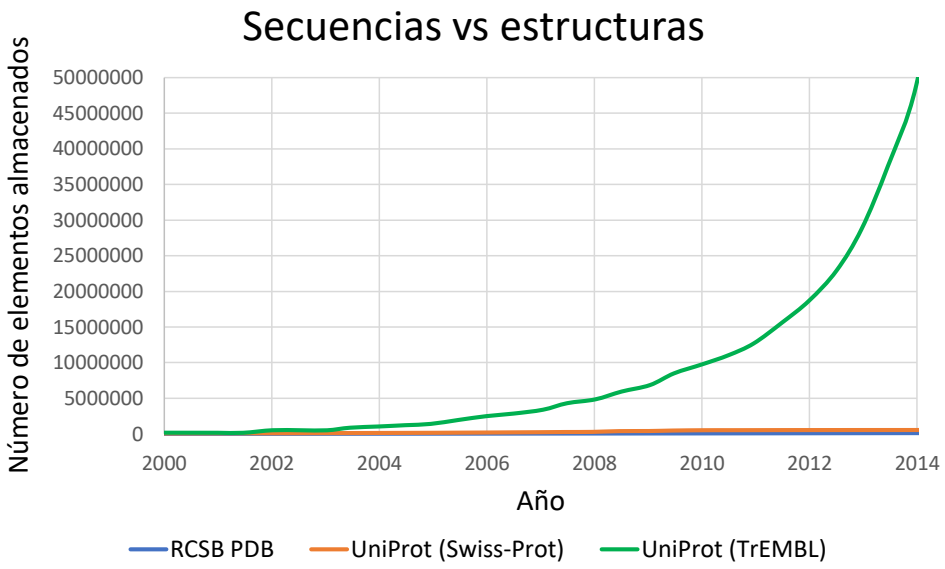

Figura 1. Comparación del crecimiento del número de secuencias almacenadas en UniProt y del número de estructuras almacenadas en el PDB. Las líneas correspondientes a RCSB PDB y UniProt (Swiss-Prot) están prácticamente superpuestas. Datos extraídos de las respectivas bases de datos.

El interés en conocer la estructura tridimensional de la proteína se debe a la relación que tiene con la función que cumple. Algunas de las aplicaciones más importantes son:

- Entender y predecir la actividad biológica que cumple una proteína: el conocimiento de la arquitectura de las proteínas es una fuente de información sobre cómo las proteínas reconocen y se enlazan con otras moléculas, cómo funcionan como enzimas[8], cómo se pliegan y cómo han evolucionado (Berg *et al.*, 2012).

- El estudio de enfermedades originadas por mutaciones genéticas como: Alzheimer, fibrosis quística, anemia de células falciformes, entre otros. Estas enfermedades surgen debido a que pequeñas modificaciones en la secuencia de una proteína puede tener un gran impacto en su estructura, lo que a su vez puede alterar su estabilidad e interacción con otras proteínas; y esto finalmente se traduce en una modificación de su actividad biológica (Soto, 2001).

- Reingeniería o rediseño de proteínas: en este caso se proponen modificaciones en ciertas regiones de la secuencia de una proteína con estructura conocida, para añadirle funcionalidad o mejorar la actividad que realiza la proteína. En el rediseño de enzimas, se busca mejorar la actividad catalítica, la (termo) estabilidad

[8] Moléculas que aceleran reacciones químicas

o solubilidad (van den Berg *et al.*, 2014). El rediseño computacional de proteínas consiste en modelar la estructura tridimensional de una proteína y predecir mutaciones de la secuencia nativa que tendrán el efecto esperado en sus propiedades bioquímicas y sus funciones (Gainza *et al.*, 2013).

- Diseño de proteínas: el cual se describe a continuación.

1.2. Diseño de Proteínas

El **diseño de proteínas** o **ingeniería de proteínas** es la construcción de nuevas moléculas de proteínas, ya sea desde cero (diseño *de novo*) o realizando modificaciones de proteínas conocidas (rediseño) (Woolfson *et al.*, 2015). El diseño *de novo* de proteínas consiste en encontrar una cadena de aminoácidos que al plegarse adoptará una estructura tridimensional deseada. Por ello, Pabo (1983) describió este problema por primera vez como el caso inverso del problema de predicción de estructura. Una característica interesante en el diseño *de novo* de proteínas es que varias secuencias de aminoácidos pueden dar lugar a una misma estructura tridimensional (Fung *et al.*, 2008; Khoury *et al.*, 2014).

Poder alcanzar los avances tecnológicos necesarios para el diseño de proteínas sería análogo a la transición de la edad de piedra a la edad de hierro (Huang *et al.*, 2016). En lugar de construir nuevas proteínas a partir de las que existen en la naturaleza, los diseñadores de proteínas pueden esmerarse en moldear nuevas moléculas para resolver problemas específicos. Las aplicaciones del diseño de proteínas son múltiples en medicina y biotecnología (Pantazes *et al.*, 2011; Huang *et al.*, 2016), por ejemplo:

- Diseño de fármacos: se pueden diseñar ligandos[9] basados en proteínas para enlazarse a estructuras objetivo como otras proteínas, virus, bacterias y células cancerígenas (Coluzza, 2017). Esto permite el tratamiento de enfermedades como Alzheimer, HIV, cáncer, fibrosis quística, entre otras (Smadbeck *et al.*, 2013; Roberts *et al.*, 2012); y también el diseño de anticuerpos (Khoury *et al.*, 2014). Más de 200 péptidos[10], proteínas, o anticuerpos terapéuticos se han comercia-

[9] El ligando es una sustancia que forma un complejo con una biomolécula (en este caso, con una proteína).
[10] Cadenas cortas de aminoácidos

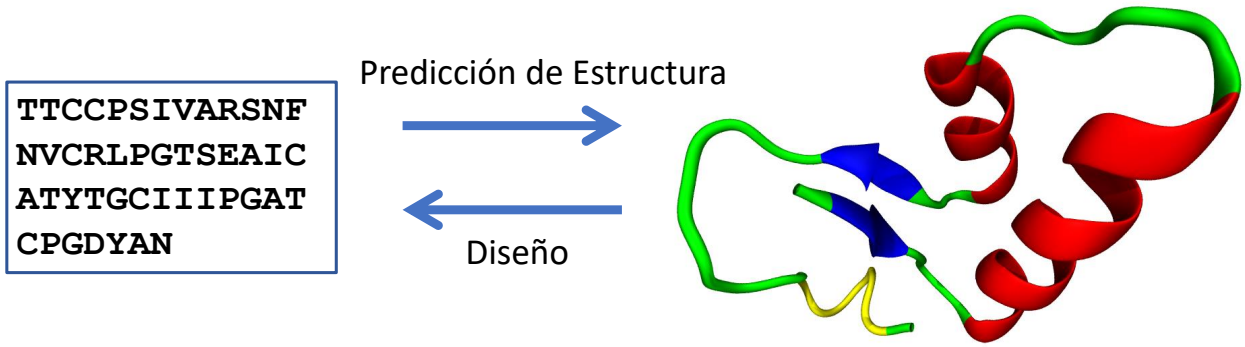

Figura 2. Predicción de estructura y diseño de proteínas. La imagen fue creada en VMD (Humphrey *et al.*, 1996) para la proteína 1CBN.

lizado al 2010 (Vlieghe *et al.*, 2010); y se espera que para el 2020 los péptidos tengan un mayor predominio dentro de los fármacos (Craik *et al.*, 2013).

- Diseño de transportadores de fármacos: el tratamiento de varias enfermedades (por ejemplo, el cáncer) requiere el uso de drogas que pueden causar efectos secundarios severos. Este impacto se puede reducir dirigiendo los fármacos específicamente a los sitios de acción contra estas enfermedades, a través de proteínas diseñadas transportadoras o nanovehículos (Chan *et al.*, 2014; Bale *et al.*, 2016).

- Diseño de biosensores: consiste en diseñar proteínas que detectan y se enlazan a pequeñas moléculas objetivo; a través de emisión de luminiscencia, cambio de color, generación de potencial eléctrico, entre otros. Por ejemplo, Griss *et al.* (2014) diseñaron sensores bioluminiscentes que permiten la medición de las concentraciones de fármacos en pacientes.

- Diseño de enzimas: además de la estructura tridimensional, se debe considerar la actividad catalítica de interés de la enzima que se desea diseñar (Baker, 2010). Se pueden diseñar enzimas que catalicen reacciones cuya existencia no se conoce en sistemas biológicos. Las aplicaciones incluyen: industria de alimentos y de detergentes, aplicaciones ambientales y fermentaciones de biocombustibles para producir etanol y butanol (Turanli-Yildiz *et al.*, 2012; Wijma y Janssen, 2013).

1.3. La predicción de estructura y el diseño de proteínas como problemas de optimización

Los métodos computacionales están basados en la hipótesis de Anfinsen (1973), que establece que las proteínas adoptan la estructura con la menor energía posible en función de su secuencia de aminoácidos. Si se dispone de una función que indique de manera precisa la energía de una estructura candidata, así como de algoritmos que puedan explorar adecuadamente el espacio de posibles estructuras y secuencias; entonces sería posible realizar completamente *in silico* la predicción de estructura y diseño de proteínas (Huang *et al.*, 2016).

Una **función de energía** modela las interacciones que ocurren en la proteína y con el medio en que se encuentra (Clote y Backofen, 2000). Estas interacciones son las interacciones de Van der Waals, enlaces de hidrógeno, enlaces iónicos y covalentes, interacción con el solvente, enlaces disulfuro, fuerzas electrostáticas, entre otras (Pearlman *et al.*, 1995; Brooks *et al.*, 2009; Gaines *et al.*, 2017). También puede incluir términos derivados del análisis estadístico de estructuras conocidas (Li *et al.*, 2013; Moult, 2005). En este caso, se la denomina **función de score**.

Por lo tanto, desde el punto de vista matemático la predicción de estructura y el diseño de proteínas son problemas de optimización (Boyd y Vandenberghe, 2004): la función objetivo es la función de score; mientras que las variables son continuas en el caso de la predicción de estructura (ángulos de torsión) y discretas en el caso del diseño de proteínas (cadenas de aminoácidos).

Para tener una idea del tamaño del espacio de búsqueda en el problema de predicción de estructura, se plantea lo que se conoce como la *paradoja de Levinthal* (Levinthal, 1969). Para ello, Levinthal consideró probar todas las conformaciones posibles que puede adoptar una proteína con 100 residuos[11], eligiendo la de menor energía como solución. Si se supone que cada residuo puede tomar tres conformaciones, el número total de estructuras posibles es 3^{100}, lo que equivale a aproximadamente 5.15×10^{47} estructuras. Suponiendo que cada estructura se evalúa en $10^{-15}s$, entonces tomaría aproximadamente 1.63×10^{25} años evaluar todas las estructuras posibles. Ciertamente las proteínas no prueban todas las combinaciones posibles durante su proceso de

[11] Los residuos se producen por la unión de aminoácidos a través de enlaces peptídicos.

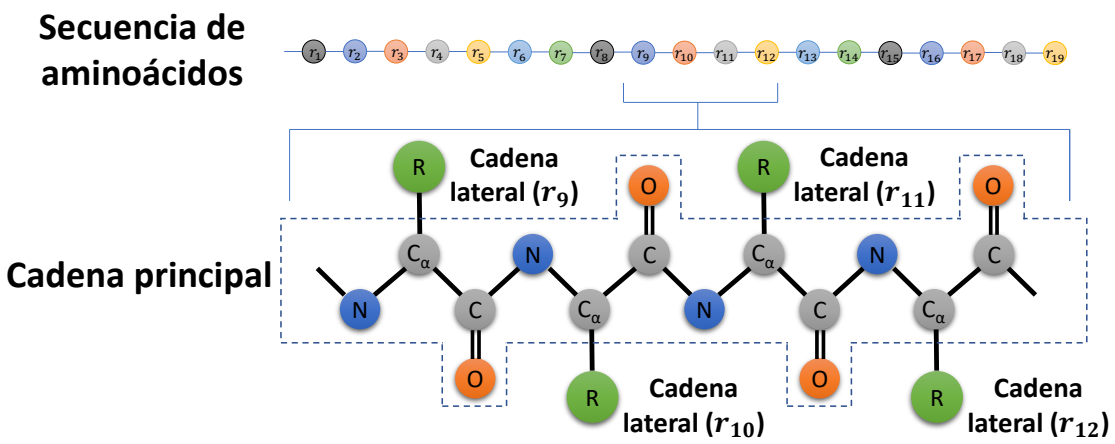

Figura 3. Cadena principal y cadena lateral de una proteína. La letra **R** agrupa a los átomos de la cadena lateral de un determinado residuo.

plegamiento (Levinthal, 1969), pero este ejercicio mental sirve para dimensionar el tamaño del espacio de búsqueda.

Algo similar se puede realizar para el diseño de proteínas. Considerando que existen 20 aminoácidos naturales, el número de secuencias posibles para una proteína de 100 residuos es 20^{100}, resultando aproximadamente 1.26×10^{130} secuencias. Si de nuevo se supone que cada secuencia se evalúa en $10^{-15}s$, se necesitarían 4.01×10^{107} años para considerar todas las secuencias posibles.

Por lo tanto, los desafíos para el enfoque computacional son: (i) que la energía de una proteína no se puede calcular con absoluta precisión, (ii) el espacio de búsqueda (de estructuras o secuencias) es muy grande y por lo tanto difícil de recorrer exhaustivamente (Huang *et al.*, 2016).

1.4. Empacamiento de la cadena lateral de proteínas

Los 20 aminoácidos naturales se diferencian entre sí por sus cadenas laterales. Los átomos que componen la estructura tridimensional de la proteína pertenecen a la cadena principal (o columna vertebral) o a la cadena lateral, como puede observarse en la Figura 3.

El presente trabajo de tesis se enfoca en un importante subproblema tanto en la predicción de estructura como en el diseño de proteínas. El **problema del empaca-**

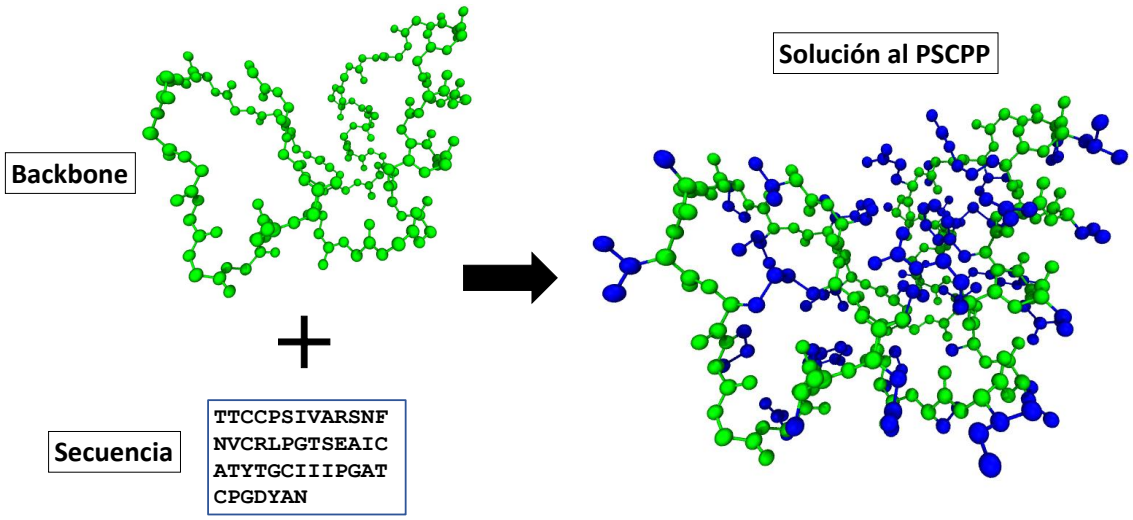

Figura 4. Problema de empacamiento de la cadena lateral de proteínas (Imagen creada con VMD (Humphrey *et al.*, 1996) para la proteína 1CBN). Los átomos de la cadena principal de la proteína están de color verde, mientras que los de la cadena lateral están en azul.

miento de la cadena lateral de proteínas (PSCPP, por sus siglas en inglés) consiste en predecir las coordenadas tridimensionales de todos los átomos de la cadena lateral de cada residuo de la proteína, basado en su secuencia de aminoácidos y las coordenadas de los átomos de la cadena principal (Colbes *et al.*, 2016). El PSCPP se representa esquemáticamente en la Figura 4.

Los métodos que intentan resolver este problema generalmente consideran que la cadena lateral de cada aminoácido puede adoptar un número finito de conformaciones, llamados **rotámeros**, que se extraen a partir de una **biblioteca de rotámeros** dada como entrada. De esta manera, el PSCPP se vuelve un problema de optimización combinatoria que consiste en seleccionar un conjunto de rotámeros (un rotámero por cada residuo) de la biblioteca para minimizar una función de score dada. Akutsu (1997) demostró que el PSCPP es un problema NP-difícil mediante la reducción del 3SAT, por lo cual no se conoce algoritmo alguno que pueda resolverlo en tiempo polinomial.

1.5. Justificación

Existen varios métodos que dan soluciones aproximadas al PSCPP, y los de mejor desempeño alcanzan una precisión de aproximadamente 87 %. Aunque han aparecido nuevas propuestas en los últimos años, no han podido mejorar esta precisión. Por otro

lado, la máxima precisión alcanzable empleando una biblioteca de rotámeros estándar es de aproximadamente 98 % (Colbes *et al.*, 2016); lo cual indica que aún hay un espacio considerable de mejora para este problema.

Surge entonces la siguiente pregunta: ¿Son las funciones de score o los algoritmos de búsqueda los principales responsables de las limitaciones de los actuales métodos para el PSCPP? Esta misma interrogante se ha planteado para la predicción de estructuras (Das, 2011) y el diseño de proteínas (Li *et al.*, 2013; Liu y Chen, 2016); y para ambos, los resultados indican que las funciones de score para estos problemas no alcanzan a modelar correctamente las interacciones que ocurren en la proteína, por lo que los esfuerzos deberían centrarse en mejorarlas y no en los algoritmos de búsqueda. Como el PSCPP es un subproblema de estos dos mencionados, generalmente las funciones de score que se emplean en él también son más simples. Por lo tanto, las funciones de score de los métodos del estado del arte para el PSCPP podrían ser la causa principal del estancamiento de los resultados actuales, y un punto importante es que no existen trabajos a la fecha que puedan responder con detalle la pregunta planteada.

Como los métodos usuales de evaluación para el PSCPP consideran sus tres componentes principales (biblioteca de rotámeros, función de score y algoritmo de búsqueda), el primer paso del presente trabajo de investigación consiste en proponer un método de evaluación que se centre únicamente en la función de score. Además de servir como evaluador de los métodos del estado del arte, este esquema podría emplearse como función evaluadora en el diseño de nuevas funciones de score para el PSCPP, con el fin de mejorar los resultados actuales.

1.6. Objetivos

1.6.1. Objetivo general

Proponer un método de evaluación de funciones de score para el PSCPP y analizar las posibles causas del estancamiento de los resultados de los métodos del estado del arte.

1.6.2. Objetivos específicos

- Determinar la máxima precisión alcanzable con bibliotecas de rotámeros estándar en métodos del estado del arte.
- Comparar el desempeño de los métodos del estado del arte para el PSCPP y determinar la máxima precisión alcanzada por los mismos.
- Proponer un método de evaluación de funciones de score para el PSCPP y emplearlo para evaluar las funciones de score de métodos del estado del arte.
- Determinar si los algoritmos basados en la búsqueda local se pueden adoptar como métodos de búsqueda estándar para el PSCPP.
- Analizar la influencia de los pesos seleccionados para los términos de la función de score empleando un enfoque multiobjetivo.
- Analizar la influencia de emplear estructuras cristalizadas como referencias para la evaluación de métodos para el PSCPP.

1.7. Contribuciones

Las principales contribuciones del presente trabajo de investigación son las siguientes:

- La comparación de cinco métodos del estado del arte para el PSCPP considerando los contactos simétricos en estructuras cristalizadas, logrando una evaluación más justa del desempeño de estos métodos. Además, se muestra la similitud de las precisiones alcanzadas por los métodos del estado del arte y la posibilidad de mejora significativa para el PSCPP al analizar la máxima precisión alcanzable de una biblioteca de rotámeros estándar.
- La propuesta del algoritmo de búsqueda local, empezando desde la estructura experimental, como método de evaluación de funciones de score para el PSCPP. También se sustenta la afirmación de que dicho método es mejor que el usado previamente para la evaluación y diseño de funciones de score (llamado *búsqueda de conformaciones de un único residuo* (Petrella *et al.*, 1998)).

- La evaluación de dos funciones de score bajo el método propuesto provee evidencia de que las funciones de score de los métodos del estado del arte para el PSCPP fallan en guiar correctamente el proceso de búsqueda hacia estructuras parecidas a la nativa, lo cual sugiere que son el principal obstáculo para la mejora de resultados.
- El análisis con un enfoque biobjetivo de los pesos de los principales términos de las dos funciones de score implementadas indicó que es irrelevante la elección de dichos pesos para mejorar los resultados actuales para el PSCPP.
- La evidencia de que el relajamiento de las estructuras cristalizadas mediante dinámica molecular, de manera a simular un ambiente más cercano al real, empeora el desempeño de las funciones de score analizadas.

1.8. Organización de la tesis

Esta tesis contiene seis capítulos y un anexo, organizados de la siguiente forma:

En el Capítulo 2 se presenta los conceptos básicos relacionados a las proteínas y se describe los métodos de determinación experimental de sus estructuras. Se presta especial atención a la cristalografía de rayos X, añadiendo definiciones que permiten entender la información contenida en los archivos PDB de las estructuras determinadas con esta técnica. Luego, se define el problema de la predicción *in silico* de estructuras de proteínas, las funciones de energía y de score, y los enfoques empleados para resolver este problema. Por último, se explica y fundamenta la representación de soluciones para este problema.

En el Capítulo 3 se define el problema estudiado en el presente trabajo de investigación: el empacamiento de la cadena lateral de proteínas (PSCPP). Debido a que la gran mayoría de los métodos propuestos para el PSCPP lo tratan como un problema de optimización combinatoria, se presenta detalladamente cada uno de los tres componentes principales que aparecen en este enfoque. Luego, se define las métricas de calidad para este problema y se describe los métodos del estado del arte seleccionados para la comparación, mostrando los resultados obtenidos. Por último, se calcula la máxima precisión alcanzable para el PSCPP.

En el Capítulo 4 se propone el método de búsqueda local para la evaluación de funciones de score, comparándolo con otro previamente propuesto en la literatura. Luego se presenta los resultados de la evaluación de las dos funciones de score implementadas en este trabajo. Una vez examinadas la biblioteca de rotámeros y la función de score, el siguiente paso consistió en evaluar la posibilidad de considerar a los algoritmos basados en la búsqueda local como estándar para el PSCPP, presentando al final del capítulo los resultados de las pruebas realizadas.

En el Capítulo 5 se analizan dos posibles razones que expliquen las limitaciones encontradas en las funciones de score evaluadas. Para ello, se plantea el PSCPP con un enfoque biobjetivo y se realiza simulaciones de dinámica molecular para relajar las estructuras cristalográficas empleadas como referencias; presentando posteriormente los resultados de los experimentos llevados a cabo.

En el Capítulo 6 se presenta un resumen de los puntos más importantes del trabajo de investigación, así como las conclusiones e ideas para trabajos futuros.

En el Anexo se provee información acerca de: (i) la obtención de las coordenadas de los átomos de la cadena lateral en función a los ángulos de torsión y parámetros de longitudes y ángulos entre enlaces, (ii) los parámetros de las funciones de score implementadas, (iii) el criterio para la vecindad entre residuos, (iv) las simulaciones de dinámica molecular, (v) las listas con los identificadores PDB de las proteínas en cada conjunto de prueba considerado, y (vi) la metodología propuesta para el diseño de funciones de score.

Capítulo 2. Marco Teórico

Este capítulo presenta los fundamentos necesarios para entender el problema considerado en este trabajo de investigación. Primero se define el área de la biología computacional y sus retos principales. Luego se introducen conceptos básicos sobre proteínas y la estructura tridimensional de las mismas, los métodos experimentales para la obtención de estructuras y su almacenamiento en el PDB. Se hace un especial énfasis en aspectos relacionados a la cristalografía de rayos X debido a que las estructuras obtenidas con esta técnica son las utilizadas para todos los experimentos de este trabajo. Finalmente, se define el problema central de este trabajo: el empacamiento de la cadena lateral de las proteínas.

2.1. Bioinformática y Biología Computacional

A comienzos de la década de los 70, la electroforesis en gel era una novedosa técnica que permitió un rápido desarrollo de la biología molecular al producir una gran cantidad de datos en bruto. Por ello fue necesario contar con algoritmos computacionales eficientes para intentar resolver el problema del análisis de los datos generados. Esto motivó el surgimiento del área interdisciplinaria conocida como **bioinformática** o **biología computacional**, que en un principio eran considerados como iguales. Clote y Backofen (2000) los definen como el área que incluye las contribuciones teóricas y prácticas de ciencias de la computación, matemáticas y biología; y trata el desarrollo de modelos matemáticos, análisis estadísticos, simulaciones por computadoras, diseño de algoritmos eficientes, sistemas de bases de datos, interfaces web, etcétera.

Con el advenimiento de las máquinas de secuenciación de siguiente generación (NGS por sus siglas en inglés), la cantidad de datos generados es aún mayor (Goodwin *et al.*, 2016). Por ello, el análisis computacional de los datos que se generan se ha vuelto cada vez más importante, y se siguen desarrollando muchas herramientas y modelos para la interpretación de estos datos biológicos.

En la última década se ha establecido una separación entre las áreas de bioinformática y biología computacional. Claverie (2000) realiza una crítica al camino tomado por los primeros bioinformáticos al establecer una analogía a las investigaciones en la física de partículas, la cual tiene dos fases. La primera fase es fenomenológica: un gran

número de eventos se almacenan en grandes bases de datos y luego se utilizan para identificar ciertas características del fenómeno estudiado. El diseño óptimo de bases de datos, de algoritmos de clasificación y de minería de datos son las áreas principales de desarrollo. El conocimiento generado en esta fase es principalmente de naturaleza estadística, y esto claramente se relaciona con la *bioinformática*.

La segunda fase persigue el descubrimiento de las reglas que explican las relaciones entre los eventos observados y sus propiedades individuales, para así explicar finalmente las distribuciones estadísticas de los eventos almacenados en las bases de datos. Lo más importante es que estas reglas permitirán la predicción de eventos que aún no se han observado. Esta fase se relaciona con la *biología computacional*.

Nair (2007) define a los bioinformáticos como los biólogos que se especializan en el uso de herramientas y sistemas computacionales para responder a problemas biológicos; mientras que los biólogos computacionales son los computólogos, matemáticos, estadísticos, e ingenieros que se especializan en desarrollar teorías, algoritmos y técnicas para tales herramientas y sistemas. Esta definición se encuentra en la misma línea que la dada por la NIH en el año 2000 (Huerta *et al.*, 2000):

- Bioinformática: consiste en la investigación, desarrollo, o aplicación de herramientas computacionales para expandir el uso de datos biológicos, médicos, conductuales, o de salud; incluyendo aquellos para adquirir, almacenar, organizar, archivar, analizar, o visualizar tales datos.
- Biología Computacional: consiste en el desarrollo y aplicación de métodos teóricos y de análisis de datos, modelado matemático y técnicas de simulación computacional para el estudio de sistemas biológicos, conductuales y sociales.

Los principales esfuerzos de investigación en la Biología Computacional incluyen el alineamiento de secuencias, el descubrimiento de genes, el ensamblado de genomas, el alineamiento estructural de proteínas, la predicción de estructuras de proteínas, la predicción de la expresión génica, las interacciones proteína-proteína y el modelado de la evolución (Smolinski *et al.*, 2009).

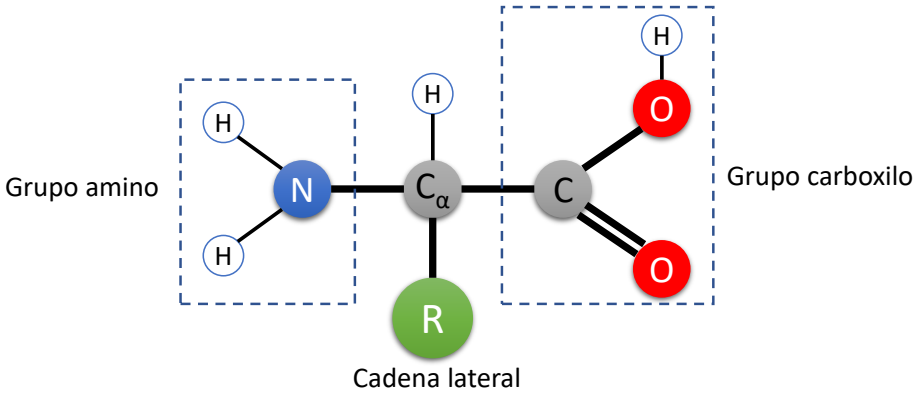

Figura 5. Composición general de un aminoácido.

2.2. Proteínas

Las proteínas son las macromoléculas más versátiles en sistemas vivos y cumplen funciones cruciales en, esencialmente, todos los procesos biológicos. Pueden funcionar como catalizadoras, transportadoras y almacenadoras de otras moléculas (el oxígeno por ejemplo), proveer soporte mecánico y protección inmune, generar movimiento, transmitir impulsos nerviosos, y controlar el crecimiento y la diferenciación celular (Berg *et al.*, 2012).

2.2.1. Aminoácidos

Las proteínas son polímeros[1] formados por la unión de unidades llamadas *aminoácidos*; por lo cual los aminoácidos son también conocidos como los bloques de construcción de una proteína. Como puede observarse en la Figura 5, un aminoácido consiste de un átomo de carbono central, denominado carbono alfa (C_α), unido a un grupo amino ($-NH_2$), a un grupo carboxilo (-COOH), a un átomo de hidrógeno y a un grupo distintivo R; siendo este último la **cadena lateral** del aminoácido.

Existen 20 tipos de aminoácidos naturales, los cuales se diferencian por la composición de sus cadenas laterales. Esto se traduce en diferencias en cuanto a tamaño, forma, carga, capacidad de formar enlaces de hidrógeno[2], hidrofobicidad y reactividad

[1] Macromoléculas formadas por la unión de unidades simples denominadas monómeros.

[2] El enlace de hidrógeno es una atracción electrostática entre dos grupos polares que ocurre cuando un átomo de hidrógeno unido mediante un enlace covalente a un átomo altamente electronegativo (como el nitrógeno o el oxígeno) experimenta el campo electrostático generado por otro átomo altamente electronegativo cercano.

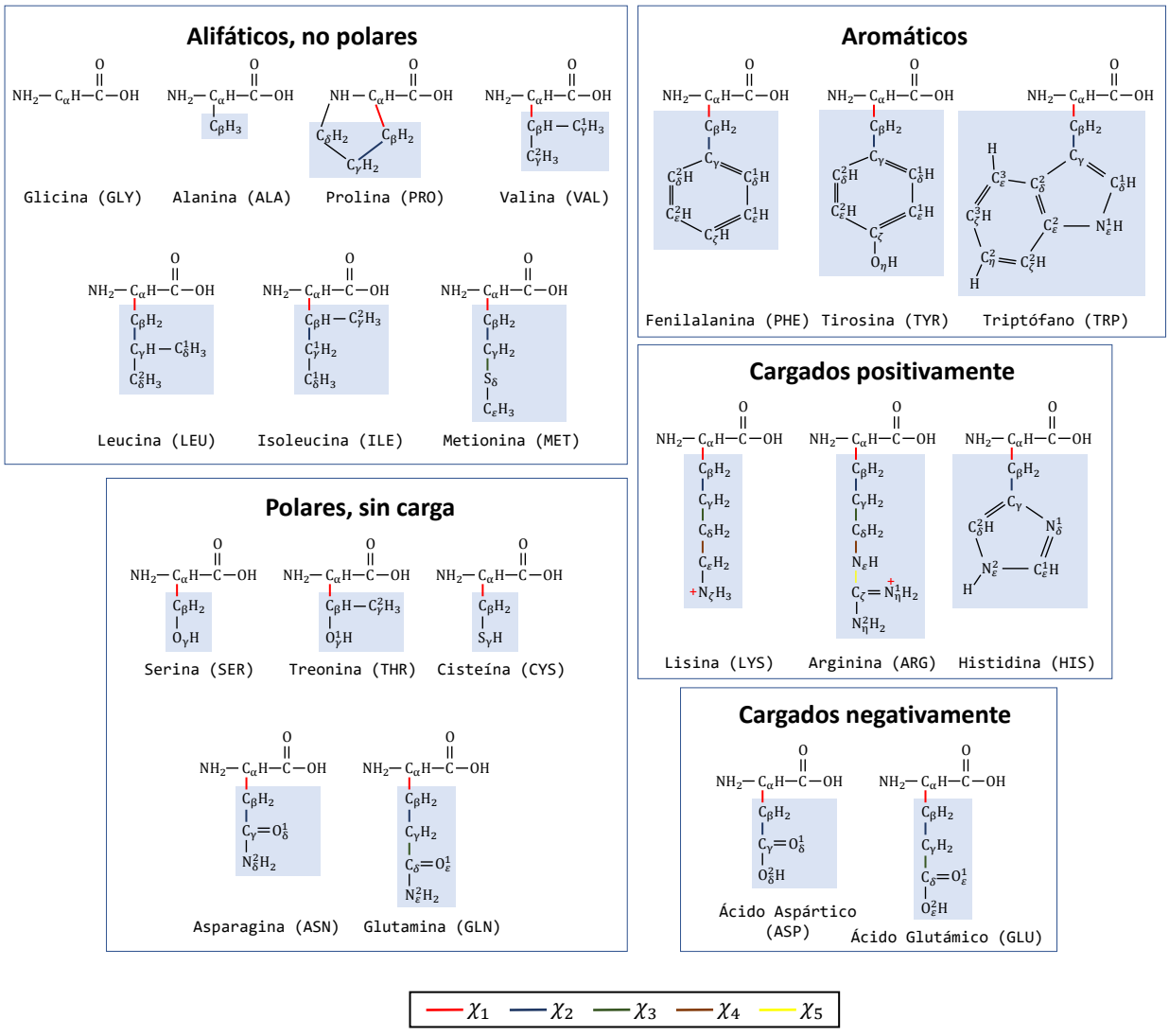

Figura 6. Aminoácidos naturales, composición y clasificación. También se muestra el identificador de tres letras para cada aminoácido. Imagen adaptada de Nelson y Cox (2004).

química. De hecho, las proteínas de todas las especies están construidas del mismo conjunto de 20 aminoácidos, con sólo algunas excepciones (Berg *et al.*, 2012). Por lo tanto, el elevado número de funciones realizadas por las proteínas resulta de la diversidad y versatilidad de estos bloques de construcción, los cuales se muestran en la Figura 6.

Los átomos de la cadena lateral de un aminoácido se etiquetan de acuerdo a la cercanía que tienen respecto al C_α, utilizando las letras griegas β, γ, δ, ϵ, ζ, y η. Si existen dos átomos a la misma distancia, se añaden números arábigos (CBN, 1970).

Los aminoácidos pueden tener identificadores de tres letras o de una sola letra.

Existen varias maneras de clasificar aminoácidos y a continuación se muestra la clasificación de Nelson y Cox (2004):

1. Grupos alifáticos[3], no-polares[4]: en este grupo los aminoácidos son no polares e hidrofóbicos[5]. Las cadenas laterales de la *alanina* (ALA, A), la *valina* (VAL, V), la *leucina* (LEU, L) y la *isoleucina* (ILE, I) tienden a agruparse dentro de la proteína, estabilizando la estructura a través de interacciones hidrofóbicas. Como la *glicina* (GLY, G) es muy pequeña, su contribución en las interacciones hidrofóbicas también lo es. La *metionina* (MET, M) realmente no es alifática ya que contiene un átomo de azufre, pero como tiene propiedades similares a las demás se la considera parte de este grupo. La *prolina* (PRO, P) tiene una estructura cíclica en la cadena lateral, lo cual le da una conformación rígida que reduce la flexibilidad estructural de la región de la proteína en donde se encuentra.

2. Grupos aromáticos[6]: La *fenilalanina* (PHE, F), la *tirosina* (TYR, Y) y el *triptófano* (TRP, W) son relativamente hidrofóbicos, por lo que pueden participar en interacciones de este tipo. El grupo hidroxilo (-OH) de la tirosina puede formar enlaces de hidrógeno y es importante para ciertas enzimas. Otra propiedad importante de este grupo es que los tres absorben la luz ultravioleta, lo cual es empleado para la caracterización de las proteínas.

3. Grupos polares, sin carga: Son más solubles en agua (es decir, más hidrofílicos), ya que pueden formar enlaces de hidrógeno con ella. Este grupo incluye a la *serina* (SER, S), la *treonina* (THR, T), la *cisteína* (CYS, C), la *asparagina* (ASN, N) y la *glutamina* (GLN, Q). Las cisteínas forman entre sí los enlaces disulfuros (volviéndose ambas hidrofóbicas), los cuales cumplen un rol especial en las estructuras de muchas proteínas.

4. Grupos cargados positivamente: los aminoácidos más hidrofílicos son aquellos cargados positiva o negativamente. En este grupo se encuentran la *lisina* (LYS,

[3] El término indica que las cadenas laterales están formadas únicamente por átomos de carbono e hidrógeno.

[4] La polaridad es una propiedad de las moléculas que representa la separación de las cargas eléctricas en la misma molécula.

[5] "Temerosos al agua": este término se emplea para indicar que estos aminoácidos tienden a interactuar entre ellos y no con el solvente (agua). El término hidrofílico expresa el caso contrario.

[6] El término indica que tienen un anillo aromático, el cual es un anillo que tiene una estructura plana estable y cuyos electrones son compartidos por todos los componentes del anillo.

K), la *arginina* (ARG, R) y la *histidina* (HIS, H). En muchas reacciones catalíticas realizadas por enzimas, la histidina facilita la reacción siendo un donador/receptor de protones.

5. Grupos cargados negativamente: en este grupo se encuentran el *ácido aspártico* (ASP, D) y el *ácido glutámico* (GLU, E), los cuales contienen un segundo grupo carboxilo (-COOH).

2.2.2. Estructura de la proteína - Niveles de representación

Dos aminoácidos pueden estar enlazados a través de un *enlace peptídico*, donde el grupo carboxilo del primer aminoácido reacciona con el grupo amino del segundo. Una serie de aminoácidos unidos por enlaces peptídicos forman una *cadena polipeptídica* (Ver Figura 7), y cada aminoácido en ella se llama *residuo*[7]. Las proteínas se forman por una o más cadenas polipeptídicas (Berg *et al.*, 2012), cuyas secuencias de residuos determinan su **estructura primaria**. Los extremos de estas cadenas se caracterizan por tener un grupo amino (por convención, a la izquierda de la cadena) y un grupo carboxilo que no son parte de algún enlace peptídico. Así, se tiene el *N terminal* como inicio y el *C terminal* como final de una cadena polipeptídica.

Una cadena polipeptídica consiste de una parte que se repite regularmente en cada residuo de aminoácido, llamada **cadena principal** o **columna vertebral**; y una parte variable, que componen las distintas cadenas laterales de los residuos, llamada simplemente **cadena lateral**. La columna vertebral está compuesta por el conjunto de átomos pertenecientes al grupo amino, grupo carboxilo y el carbono-α de todos los residuos de la cadena. En general sólo se consideran los átomos pesados en una proteína, dejando afuera así a los átomos de hidrógeno. Entonces, los átomos de cada residuo de la cadena polipeptídica que pertenecen a la cadena principal son: el nitrógeno (N), el carbono alfa (C_α), el carbono carbonilo (C) y el oxígeno carbonilo (O). Por la composición de la cadena principal, esta región tiene un alto potencial para formar enlaces de hidrógeno.

Una cadena polipeptídica puede adoptar ciertas estructuras regulares, siendo las principales las estructuras periódicas llamadas *hélices* α y *láminas* β. Una hélice α es

[7] Como al formar un enlace peptídico se libera una molécula de agua, entonces quedan *residuos* de aminoácidos en cada lado.

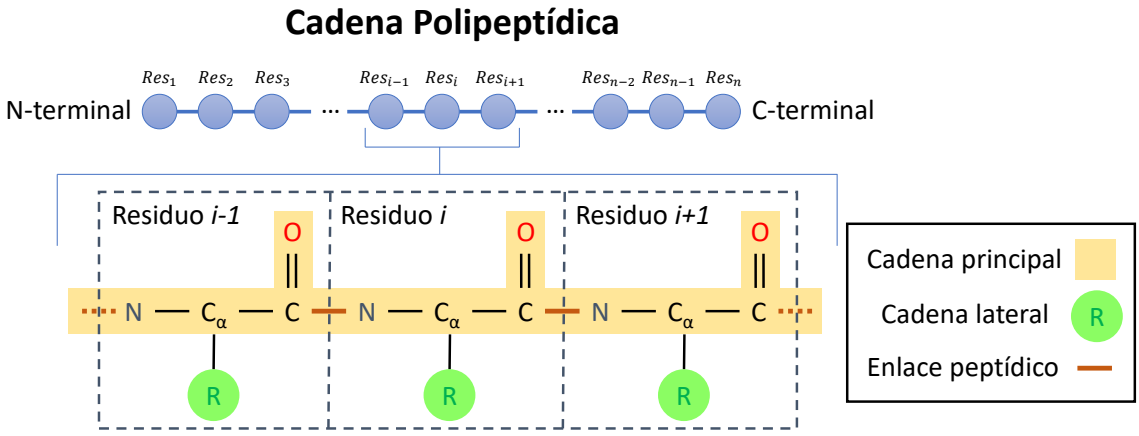

Figura 7. Composición de una cadena polipeptídica, formación de enlaces peptídicos y diferenciación entre cadena principal y lateral. En la cadena principal de esta figura sólo se muestran los átomos pesados.

generada por el apilamiento de residuos en una hélice, formando un cilindro fijo. Una lámina *β* es una estructura donde parte de la cadena de residuos se apilan de una manera lineal. Como la proteína tiene una dirección (del término N al C), se pueden distinguir las hojas paralelas (las partes tienen la misma dirección) y antiparalelas (dirección opuesta).

También existen estructuras irregulares como los *lazos*, que conectan a las dos anteriores; y aunque no son periódicas están bien definidas y contribuyen junto con las *hélices α* y *láminas β* a formar la estructura final de la proteína. Estas tres estructuras están formadas por un patrón regular de enlaces de hidrógeno entre los átomos de la cadena principal de residuos cercanos dentro de la secuencia de la cadena polipeptídica y dichas estructuras se conocen como elementos de la **estructura secundaria** de la proteína.

La estructura tridimensional compacta y asimétrica que adopta individualmente cada cadena polipeptídica se conoce como **estructura terciaria**. Desde este nivel de descripción es útil la siguiente clasificación de proteínas de acuerdo a sus características estructurales y de solubilidad (Nelson y Cox, 2004):

1. Proteínas fibrosas: cumplen varias funciones estructurales, y consisten en cadenas polipeptídicas dispuestas en múltiples repeticiones de elementos simples con estructura secundaria; por lo que resultan en estructuras largas de un solo tipo de estructura secundaria. Todas las proteínas fibrosas son insolubles en agua, lo

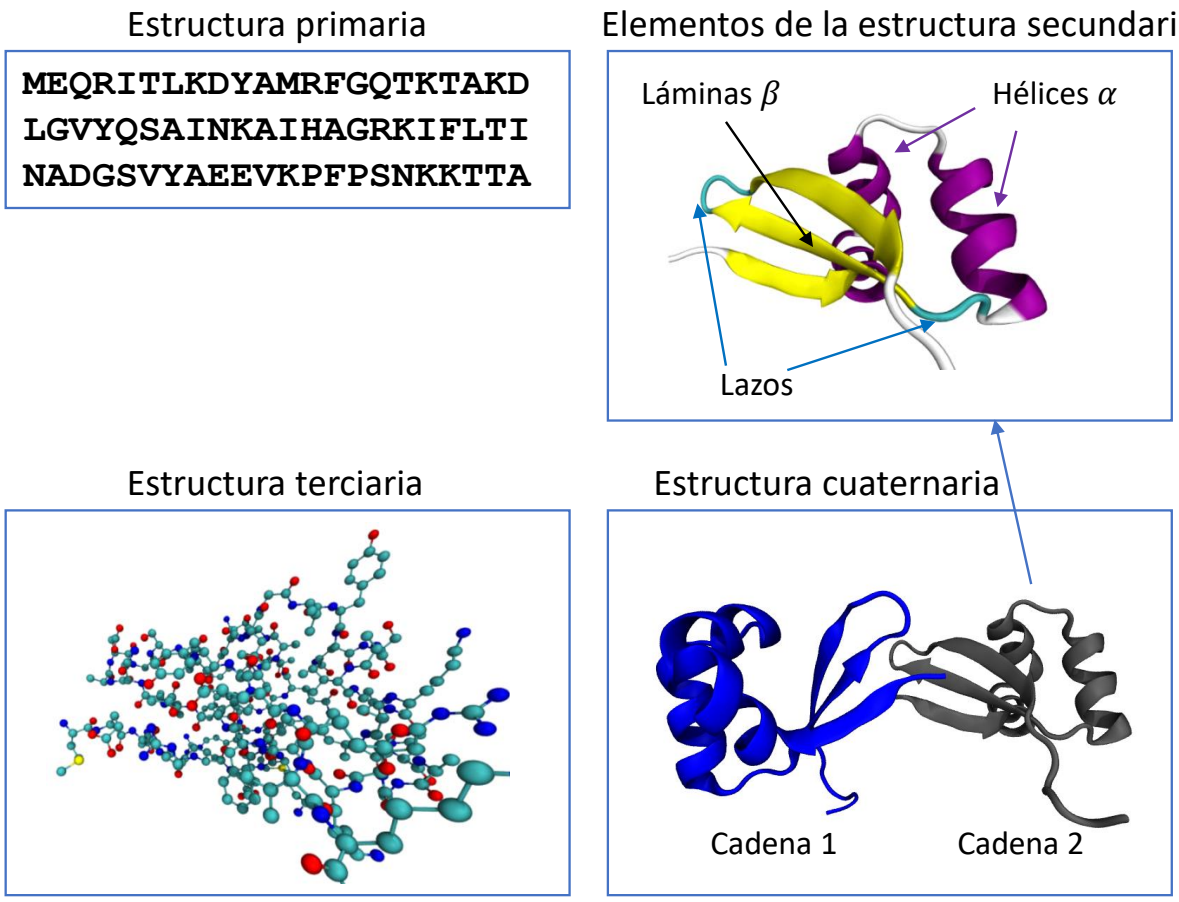

Figura 8. Niveles de representación estructural de la proteína 5CRO. Las imágenes fueron creadas con el programa VMD (Humphrey *et al.*, 1996). Existen varias formas de visualizar estructuras tridimensionales; en esta imagen se adoptaron los estilos “Cartoon” y “CPK”.

cual se debe a la alta concentración de residuos hidrofóbicos en el interior y en la superficie de la proteína. Algunos ejemplos son la α-queratina y el colágeno.

2. Proteínas globulares: incluyen a las enzimas y a las proteínas de transporte, motoras, reguladoras, inmunoglobinas y otras. Las proteínas globulares tienen cadenas polipeptídicas que adoptan una forma esférica o globular y tienen un patrón de plegado más complejo que las fibrosas. Las proteínas de este tipo generalmente contienen varios tipos de estructura secundaria y son más solubles en agua. La mioglobina[8] fue la primera proteína de este tipo cuya estructura fue determinada experimentalmente.

3. Proteínas de membrana: son proteínas que interactúan o se encuentran en membranas biológicas. Como las membranas están constituidas principalmente de

[8] Proteína que almacena oxígeno.

cadenas hidrofóbicas de alcanos[9], los residuos superficiales de las proteínas de membrana son también hidrofóbicos con el fin de lograr la interacción. Las proteínas de este tipo funcionan como receptoras o como transportadoras a través de la membrana, por lo que pueden cruzar la misma o sólo interactuar con su parte interna o externa. Un ejemplo es la proteína RAS, que es importante en rutas de transmisión de señales celulares.

Las estructuras de las proteínas solubles en agua tienen dos características en común: (i) su interior está formado por residuos de aminoácidos hidrofóbicos, (ii) su superficie está formada principalmente por residuos de aminoácidos hidrofílicos que interactúan con el solvente, denominados residuos *expuestos*. Por lo tanto, las interacciones hidrofóbicas de los residuos internos (conocidos también como residuos *enterrados*), son la fuerza principal para la formación de la estructura terciaria de este tipo de proteínas. A este proceso de formación de la estructura tridimensional se lo conoce como *plegamiento de la proteína*. En las proteínas de membrana ocurre lo contrario: los residuos hidrofóbicos están en la superficie e interactúan con el medio hidrofóbico en que se encuentran, mientras que los residuos hidrofílicos se encuentran blindados dentro de la proteína (Berg *et al.*, 2012).

Las proteínas que consisten de más de una cadena polipeptídica presentan una **estructura cuaternaria**, donde cada cadena se llama *subunidad*. Las estructuras cuaternarias pueden ser tan simples como dos subunidades idénticas (ver ejemplo de la Figura 8), o complejas como docenas de subunidades distintas. En la mayoría de los casos, las subunidades se mantienen juntas mediante enlaces no covalentes (Berg *et al.*, 2012).

2.2.3. Determinación experimental de estructuras de proteínas

Actualmente, la determinación de la estructuras tridimensionales de proteínas de interés se logra a través de métodos experimentales, siendo los más utilizados los siguientes:

1. *Cristalografía de rayos X*: alrededor del 89 %[10] de las estructuras conocidas se

[9] Compuestos que sólo contienen átomos de carbono e hidrógeno.
[10] www.rcsb.org/pdb/statistics/holdings.do

obtuvieron por este método. En este método las proteínas se cristalizan, formando un configuración en donde las proteínas están orientadas de una manera fija y repetitiva entre sí. Un haz de rayos X se dirige al cristal, y la difracción de este haz por los átomos de la proteína se transforma a un mapa de densidad de electrones. Este mapa se usa finalmente para inferir las posiciones de los átomos de la proteína. La ventaja de este método es la obtención de estructuras de alta resolución, pero la desventaja principal está en que la cristalización conduce a un ambiente no-nativo de la proteína y puede afectar su estructura. Además, ciertas proteínas son difíciles de cristalizar (Berg *et al.*, 2012).

2. *Resonancia magnética nuclear* (*NMR* por sus sigas en inglés): Aproximadamente el 9 %[11] de las estructuras almacenadas se determinaron con esta técnica. El NMR se basa en el hecho de que ciertos núcleos atómicos son intrínsecamente magnéticos. Por ello, la proteína en solución se expone a un campo magnético externo y a una radiación electromagnética, ocasionando que ciertos átomos de la proteína (por ejemplo: el hidrógeno) se exciten y emitan una radiación electromagnética en función a su ambiente circundante. Esta información permite determinar la ubicación de cada átomo en relación a los demás dentro de la estructura, y se generan múltiples modelos para representar los posibles estados tridimensionales de la proteína. Este método no logra generalmente alcanzar el mismo nivel de resolución que la cristalografía de rayos X y está limitado a proteínas pequeñas, pero provee información adicional acerca de la dinámica de una proteína en solución (Berg *et al.*, 2012).

3. *Crio-Microscopía Electrónica* o *Nanoscopía* (*Cryo-EM por sus siglas en inglés*): A pesar de que sólo un poco más del 1 % de las estructuras almacenadas se han obtenido con este método, su uso para el estudio de estructuras de proteínas se está extendiendo cada vez más (ver Figura 9). En este método básicamente se utiliza un microscopio electrónico para obtener imágenes de la proteína y determinar así su estructura tridimensional. Debido a que las proteínas son sensibles al haz de electrones aplicado y pueden degradarse, las mismas se congelan para atenuar el impacto. Con los recientes avances se está alcanzando la resolución lograda por la cristalografía de rayos X, aunque no se necesita una cristalización y permite el análisis de proteínas de gran tamaño o complejos de proteínas (Bai

[11] www.rcsb.org/pdb/statistics/holdings.do

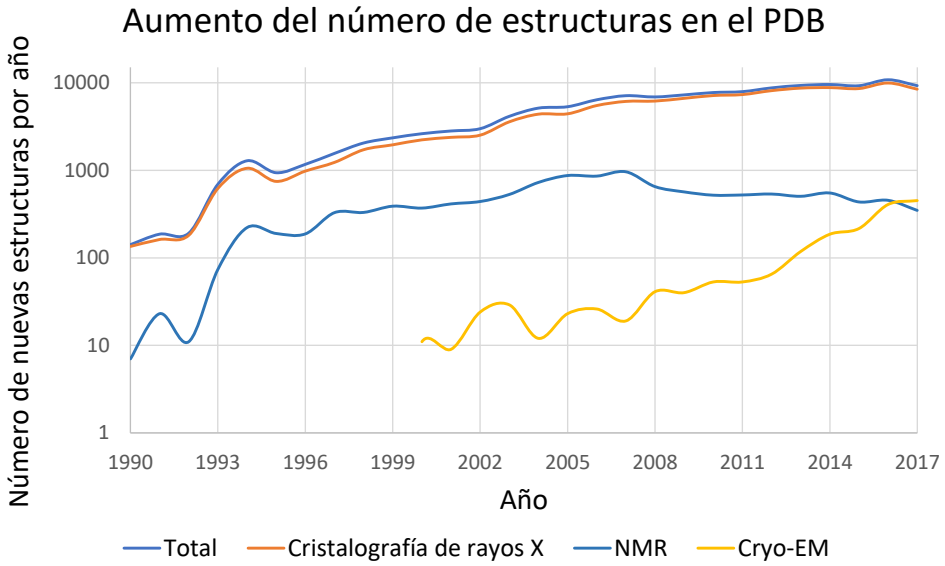

Figura 9. Número de nuevas estructuras depositadas en el banco de datos de proteínas (PDB por sus siglas en inglés) por año. El PDB almacena estructuras obtenidas mediante métodos experimentales. Puede notarse el aumento del uso de la Crio-Microscopía Electrónica en los últimos años. Los datos fueron obtenidos de las estadísticas en rcsb.org.

et al., 2015).

2.2.3.1. Aspectos relacionados a la cristalografía de rayos X

El proceso de generación de la estructura de una proteína mediante la cristalografía de rayos X puede entenderse a un nivel elemental considerando el funcionamiento de un microscopio óptico. En él, la luz de una fuente puntual se enfoca en un objeto. El objeto dispersa las ondas de luz, y estas ondas dispersas se recombinan mediante un arreglo de lentes para generar una imagen ampliada del objeto en cuestión. El objeto más pequeño cuya estructura puede determinarse por este sistema está dado por la longitud de onda de la luz (en este caso la luz visible), y a esta longitud de onda se la conoce como la *resolución* del microscopio. Los objetos más pequeños que la mitad de la longitud de onda de la luz incidente no se pueden reconocer (Nelson y Cox, 2004).

Para la determinación de estructuras del tamaño de las proteínas es necesaria una longitud de onda menor a la de la luz visible. De hecho, los rayos X proveen la mejor resolución para la determinación de estructuras moleculares, ya que sus longitudes de onda aproximadamente corresponden a los tamaños de los enlaces covalentes (Berg

et al., 2012). Los tres componentes del método de cristalografía de rayos X son: la proteína cristalizada, una fuente de rayos X y un detector.

Primero, las proteínas deben prepararse para obtener el cristal. Al añadir lentamente sulfato de amonio u otra sal a una solución concentrada de proteínas, se reduce su solubilidad y se favorece así la formación de cristales altamente ordenados. Un aspecto importante es que no existe una sola proteína dentro del cristal, sino que se tiene un arreglo fijo y orientado de múltiples copias de la misma proteína (es decir, un complejo de proteínas). La existencia de múltiples *copias* de una proteína a su alrededor puede provocar la aparición de interacciones entre proteínas, lo que sumado a la variación de su solubilidad puede ocasionar la variación de la estructura nativa de la proteína (Nelson y Cox, 2004).

Cuando un haz de rayos X es dirigido a la proteína cristalizada, la mayor parte del mismo pasa directamente a través del cristal, mientras que una pequeña parte es dispersa en varias direcciones. El detector capta estos rayos X dispersos, con lo cual se forma una *fotografía* de rayos X que consiste en un arreglo regular de puntos denominados *reflexiones*, donde cada átomo de la molécula realiza una contribución a cada punto. La parte izquierda de la Figura 10 presenta los componentes principales de la cristalografía de rayos X.

La diferencia principal con el microscopio óptico radica en que no existen lentes apropiados para formar la imagen ampliada de la molécula a partir del patrón de dispersión de los rayos X. En lugar de ello, la aplicación de la transformada de Fourier a las *amplitudes* y *fases* de cada reflexión observada[12] reconstruye un *mapa de densidad electrónica* de la proteína. Este mapa es una representación gráfica de tres dimensiones que indica dónde los electrones están más densamente localizados, y es utilizado para determinar las posiciones de los átomos en la molécula cristalizada, como puede observarse en la parte derecha de la Figura 10.

Por lo tanto, la calidad o resolución del mapa de densidad electrónica es crítica para su interpretación. La **resolución** se mide en Åmströng (Å) y se puede definir como el mínimo espacio *d* entre los planos de la red cristalina que provee una difracción medible de los rayos X, y la calidad del cristal empleado para el experimento también

[12] Estas amplitudes y fases se conocen como *factores de estructura*.

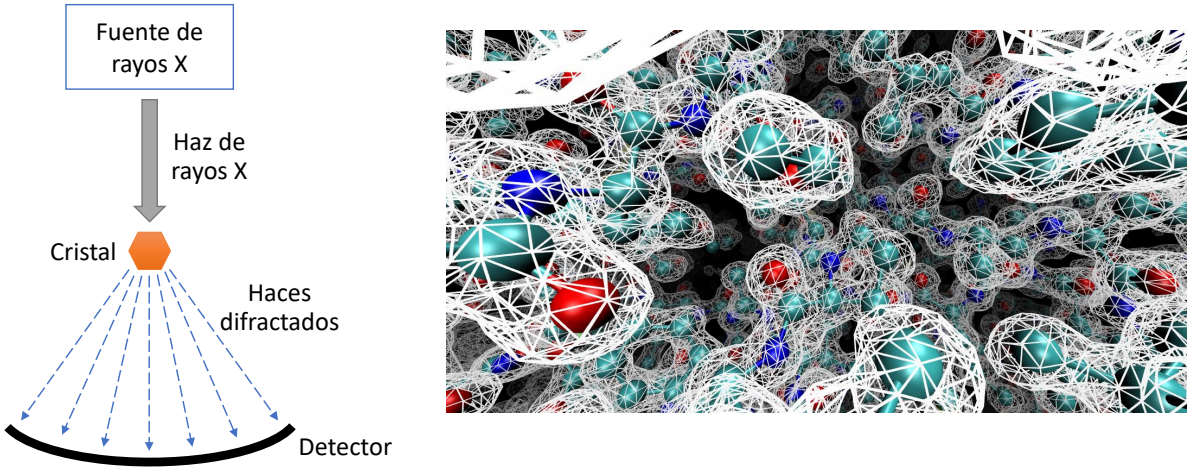

Figura 10. *Izquierda*: Componentes de la cristalografía de rayos X. Una fuente de rayos X genera un haz, el cual es difractado por el cristal. El patrón de difracción resultante es recogido por el colector (Adaptación de la imagen en (Berg *et al.*, 2012)). *Derecha*: Las líneas blancas representan el mapa de densidad electrónica obtenido a partir del patrón de difracción, y sirven para obtener el modelo con la ubicación de cada átomo de la proteína estudiada. La imagen fue generada con VMD (Humphrey *et al.*, 1996) para la proteína 1A7S.

afecta este valor. La distancia *d* define el nivel de detalle de la estructura obtenida, o dicho de otra manera, la mínima distancia entre características estructurales que pueden distinguirse en los mapas de densidad electrónica. Cuanto mayor sea la resolución (ésto es, cuanto menor sea *d*) mejor es la calidad de la imagen obtenida, pues existen más reflexiones independientes disponibles para determinar la imagen (Wlodawer *et al.*, 2008), y mayor es la cantidad de detalles que pueden observarse en la proteína estudiada. Por lo tanto, la resolución también se puede determinar por medio del número de intensidades de dispersión empleadas en la transformada de Fourier.

Los términos usualmente empleados para referirse a los niveles de resolución son “bajo”, “mediano”, “alto” y “atómico”; como puede verse en la Figura 11. Las estructuras cristalizadas con menor resolución que se han publicado tienen una resolución de aproximadamente 6 Å, lo cual es suficiente para proveer una idea muy básica de la forma de la macromolécula. En la actualidad se tienen resoluciones mejores que 2 Å, el cual es el valor de referencia en varios conjuntos de prueba para el PSCPP. El valor de 1.5 Å corresponde al tamaño típico de enlaces covalentes entre átomos de carbono; y a 1.2 Å se alcanza la resolución atómica, que corresponde a la distancia de interacción más corta entre átomos que no sean de hidrógeno (Wlodawer *et al.*, 2008).

Como se mencionó anteriormente, el resultado del experimento mediante la difrac-

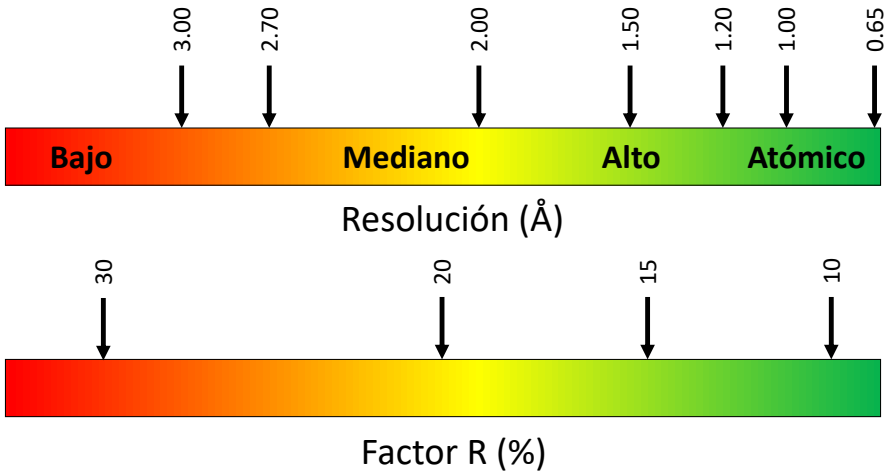

Figura 11. Resolución y factor R como medidas de calidad de modelos cristalográficos de estructuras macromoleculares. Adaptación de la imagen en (Wlodawer *et al.*, 2008).

ción del haz de rayos X es el mapa de densidad electrónica dentro del cristal; donde el modelo, consistente de las coordenadas de cada átomo de la proteína, es únicamente una interpretación de la densidad electrónica. Este modelo se *refina* mediante la variación de todos los parámetros del modelo para lograr la mejor concordancia entre las amplitudes de reflexión observadas (F_{obs}) y las calculadas a partir del modelo propuesto (F_{calc}). Esta concordancia se mide por el factor residual o cristalográfico, más conocido como **factor R**, que se define como:

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}} \times 100\,\% \qquad (1)$$

El factor R combina el error inherente en los datos experimentales y la desviación del modelo de la realidad (Wlodawer *et al.*, 2008). Se espera que las estructuras macromoleculares bien refinadas tengan un factor R menor a 20 %, ya que para estos valores generalmente se tiene la topología correcta (Morris *et al.*, 1992).

En las estructuras obtenidas por cristalografía de rayos X aparecen los conceptos de *ocupación* y *factor de temperatura*[13]. En el cristal existen múltiples copias de una misma macromolécula y puede darse el caso de que existan pequeñas diferencias de conformaciones entre ciertas regiones de cada una de estas moléculas. Al construir los modelos atómicos de estas regiones, los cristalógrafos pueden usar la **ocupación**

[13] https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/dealing-with-coordinates

para estimar la cantidad de cada conformación observada en el cristal. Para la mayoría de los átomos, la ocupación tiene un valor de 1, y este valor indica que la posición determinada para ese átomo se encontró en todas las moléculas del cristal. En cambio, si por ejemplo un ion metálico se une sólo a la mitad de las moléculas en el cristal, el cristalógrafo observará una imagen débil del ion en el mapa de densidad electrónica y puede asignarle una ocupación de 0.5. Este ion también puede afectar la ubicación de ciertos átomos vecinos que forman parte de la macromolécula estudiada, por lo cual estos átomos pueden tener diferentes conformaciones con valores de ocupación menores a 1. Las ocupaciones se usan también comúnmente para identificar átomos de la cadena lateral que son observados en múltiples conformaciones. El valor de ocupación indica la fracción de moléculas del cristal que tienen una cierta conformación, por lo que la suma de las ocupaciones para las conformaciones de un átomo de la macromolécula debe ser igual a 1. Cuando se desea seleccionar una única conformación para cierto átomo, generalmente se toma el de mayor ocupación.

Si de alguna manera se pudiese mantener un átomo rígidamente en cierta posición, podría observarse su distribución electrónica en una situación ideal. La imagen sería densa en el centro y la densidad disminuiría al alejarse del mismo. Pero al observar la distribución electrónica en los datos experimentales, usualmente los electrones tienen una distribución más amplia a la ideal. Esto puede deberse a la vibración de los átomos o a las diferencias entre las distintas moléculas en el cristal. La densidad electrónica observada es el promedio de todas estas pequeñas variaciones, formándose así una imagen ligeramente “embarrada” de la molécula. Estas variaciones, y la densidad electrónica embarrada resultante, se incorporan al modelo mediante el **factor B** o **factor de temperatura** para cada uno de los átomos. Básicamente, el factor de temperatura es una medida de la confianza en la ubicación de cada átomo. Los valores inferiores a 10 crean un modelo muy nítido del átomo, lo cual indica que éste no se mueve mucho y está en la misma posición en todas las moléculas del cristal. Los valores mayores a 50 indican que el átomo se mueve tanto que apenas se puede observar. Éste suele ser el caso de los átomos en la superficie de la proteína, donde las cadenas laterales largas tienen la libertad de moverse en el agua circundante.

Otro aspecto muy importante sobre las estructuras obtenidas mediante cristalogra-

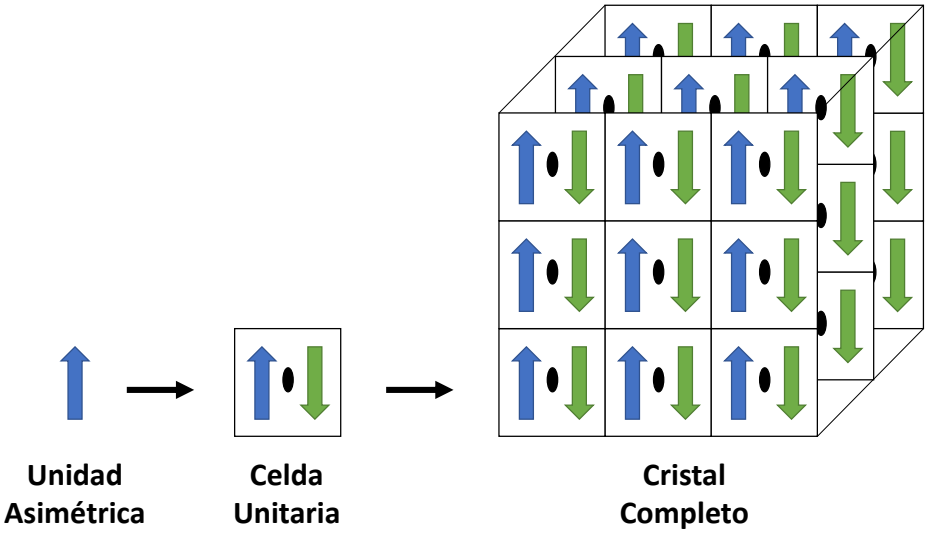

Figura 12. En este ejemplo, la unidad asimétrica se rota 180° alrededor de un eje de simetría (el óvalo negro) para producir una segunda copia, y juntas generan la celda unitaria. Ésta a su vez se repite y se traslada en las tres direcciones para generar el cristal tridimensional. Adaptación de la imagen en https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies.

fía de rayos X es la diferenciación entre unidad asimétrica y ensamblado biológico[14] (Berman *et al.*, 2000). El **ensamblado biológico** (o unidad biológica) es el ensamblado macromolecular que se ha demostrado o se cree que es la forma funcional de la molécula. Por ejemplo, la forma funcional de la hemoglobina[15] contiene cuatro cadenas.

Por otro lado, la **unidad asimétrica** es la porción más pequeña de la estructura del cristal a la cual se puede aplicar operaciones de simetría (rotaciones y traslaciones) para generar la *celda unitaria* completa, con la cual puede generarse el cristal completo mediante operaciones de traslación (ver Figura 12). Ciertas estructuras almacenadas en el PDB reportan únicamente las posiciones de los átomos que pertenecen a una unidad asimétrica, aunque indican las operaciones de rotación y traslación que se deben realizar con las copias necesarias para obtener la celda unitaria completa.

Como el cristal se forma mediante múltiples copias de la celda unitaria, puede darse el caso de que los residuos que se encuentran en una celda unitaria interactúen con otros residuos de otra celda unitaria. Ésta es una de las razones por la cual la proteína obtenida mediante cristalización por rayos X puede no estar en su conformación

[14] https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies
[15] Proteína en los glóbulos rojos que transporta oxígeno.

nativa, debido a estas interacciones proteína-proteína. El término **contacto simétrico** se emplea para indicar la interacción existente entre residuos de diferentes celdas unitarias.

El cristalógrafo usa la estructura asimétrica para refinar las coordenadas de la estructura con los datos experimentales, y no necesariamente representan el ensamblado biológico de la proteína. Una unidad asimétrica del cristal puede contener el ensamblado biológico, parte del mismo, o múltiples copias de él[16]. Dependiendo de las condiciones de cristalización y el empacamiento local de las proteínas, puede ocurrir uno de los siguientes escenarios:

- Las copias de la macromolécula dentro de la celda unitaria tienen conformaciones idénticas u ocupan posiciones relacionadas simétricamente. En este caso, el ensamblado biológico puede estar compuesto de una copia de la macromolécula, o por dos o más copias relacionadas simétricamente que juntas forman el ensamblado biológico.
- Las copias de la macromolécula tienen conformaciones ligeramente diferentes y ocupan posiciones únicas en la unidad asimétrica del cristal. Como resultado, cada una de estas conformaciones diferentes puede corresponder a ensamblados biológicos estructuralmente similares pero no idénticos.

2.2.4. Almacenamiento de estructuras en el PDB

El banco de datos de proteínas (*Protein Data Bank* - PDB) es un repositorio de estructuras tridimensionales determinadas experimentalmente de macromoléculas biológicas, las cuales están disponibles públicamente (Berman *et al.*, 2000). Cada proteína almacenada en el PDB posee un archivo individual en el cual se encuentran los datos sobre las coordenadas atómicas (en Åmströngs), factores de estructura cristalográfica, datos experimentales de la resonancia magnética nuclear y otros. Además de las coordenadas de cada átomo, se incluye información sobre el nombre de la molécula, la estructura primaria y secundaria, referencias en bases de datos de secuencias, información sobre ligandos[17], detalles acerca de la recolección de datos y la solución

[16] Existen estructuras de la hemoglobina en el PDB para cada caso (2HHB, 1OUT y 1HV4).
[17] Sustancia que forma un complejo con una biomolécula (en este caso, con una proteína).

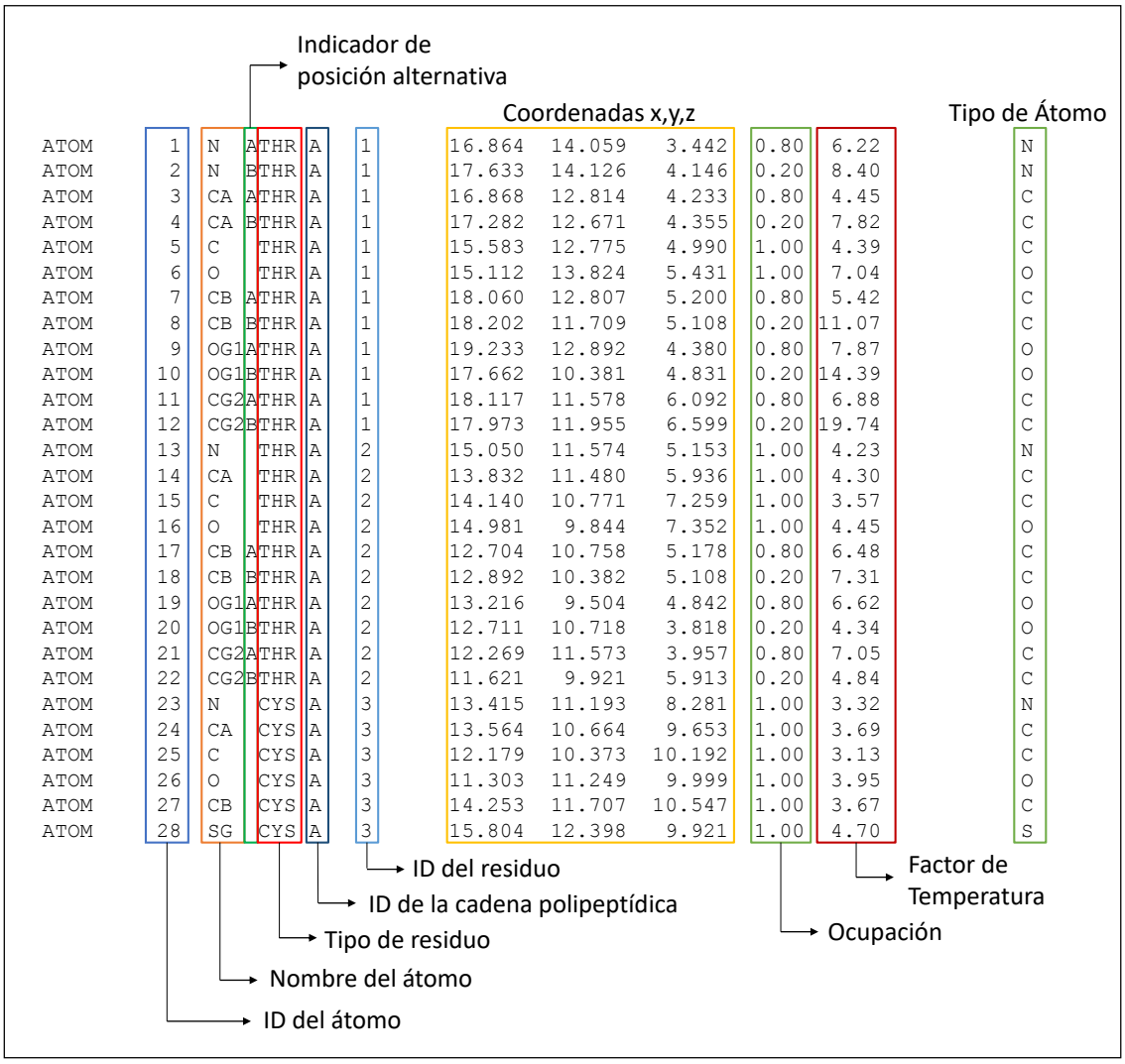

Figura 13. Ejemplo de la presentación en un archivo PDB de los datos sobre los átomos de la proteína.

de la estructura, y citas bibliográficas[18].

Cada estructura tridimensional almacenada en el PDB tiene un identificador único de cuatro caracteres, donde el primero es un número y el resto son caracteres alfanuméricos. Puede darse el caso donde una misma proteína tenga más de una estructura experimental almacenada en el PDB. Por ejemplo, para la hemoglobina se tienen las estructuras con identificadores 2HHB, 1OUT y 1HV4.

Los archivos en el PDB pueden tener estructuras incompletas, donde ciertos residuos aparecen en la estructura primaria (SEQRES) pero sus átomos no poseen coordenadas tridimensionales. Para estos casos se encuentra el “REMARK 465” que se

[18]Ver http://www.wwpdb.org/documentation/file-format para más detalles

incluye en el encabezado de los archivos PDB e indica los residuos que faltan.

En la Figura 13 se puede observar el formato general para presentar la información relativa a los átomos de la proteína.

2.3. Predicción *in silico* de estructuras de proteínas

Anfinsen (1973) realizó una serie de experimentos para apoyar la hipótesis principal para el plegamiento de una proteína a su estructura nativa. El experimento más sobresaliente consistió en el reordenamiento de la ribonucleasa[19]. Para ello, Anfinsen *desnaturalizó*[20] la ribonucleasa mediante un reactivo (*urea*), con lo cual también la enzima perdió su actividad catalítica. Anfinsen encontró que la proteína recuperó su actividad enzimática al remover el reactivo, con lo cual concluyó que la enzima recuperó su estructura original.

Anfinsen entonces propuso lo que se conoce como **hipótesis termodinámica**, la cual establece que la estructura tridimensional nativa de la proteína en su ambiente fisiológico normal (solvente, pH, presencia de otros componentes como iones metálicos o grupos prostéticos, temperatura y otros) es la de menor energía considerando el sistema completo. Esto es, la conformación nativa la determinan las interacciones atómicas (y por lo tanto, la secuencia de aminoácidos) en un ambiente dado (Anfinsen, 1973).

Los enlaces de hidrógeno e iónicos, el efecto hidrofóbico y las interacciones de Van der Waals[21] son individualmente débiles, pero colectivamente tienen una influencia significativa en la estructura tridimensional de las proteínas (Nelson y Cox, 2004). Estas características se deben considerar en las funciones de energía.

[19] Enzima que cataliza la degradación del ARN en moléculas más pequeñas.
[20] La desnaturalización de una proteína consiste en modificar su estructura nativa.
[21] Será definido con mayor detalle más adelante, pero las fuerzas de Van der Waals básicamente son interacciones (de atracción o repulsión) entre átomos no relacionados mediante enlaces químicos, y sus valores dependen de la distancia entre los átomos considerados.

2.3.1. Funciones de energía (o de score)

La **función de energía** es un modelo que aproxima la energía resultante de las interacciones entre los átomos de una proteína y con el medio en que se encuentra dicha proteína; y su uso se extiende al estudio de moléculas de interés en bioquímica y química orgánica (Clote y Backofen, 2000). Existen dos tipos principales de funciones de energía (Lazaridis y Karplus, 2000; Li *et al.*, 2013): (i) las *basadas en la física* (*physical effective energy function - PEEF*) y (ii) las *basadas en el conocimiento* (*statistical effective energy function - SEEF*).

El término **función de score** se introduce para generalizar las funciones que incluyen términos basados en el conocimiento. Las funciones de score deben ser lo suficientemente precisas para capturar las características más importantes de las interacciones atómicas; y al mismo tiempo, deben tener una implementación computacional rápida (Gordon *et al.*, 1999; Boas y Harbury, 2007; Suarez y Jaramillo, 2009). Una de las propiedades que se espera de una función de score efectiva es su habilidad de reconocer a la estructura nativa, y poder usarse entonces en la predicción de estructuras y el diseño de proteínas (Khoury *et al.*, 2014).

2.3.1.1. Funciones de energía basadas en la física

Este tipo de función de energía se conoce también como función basada en los *primeros principios*. Idealmente consiste en la función de energía *real*, derivada por principios físico-químicos que abarcan el entendimiento de las estructuras químicas y las interacciones a nivel atómico, donde generalmente son necesarios ciertos cálculos de mecánica cuántica. A pesar de los progresos en este tipo de función de energía, las aproximaciones requeridas para modelar de forma correcta los computacionalmente complejos efectos electrostáticos y del solvente, siguen siendo el principal reto. Ponder y Richards (1987) señalan que el éxito de la función de energía basada en los primeros principios permitirá asegurar el entendimiento de la estructura de las proteínas. Dos ejemplos clásicos de funciones de energía basadas en la física son AMBER (Cornell *et al.*, 1995) y CHARMM (Brooks *et al.*, 2009).

2.3.1.2. Funciones de score basadas en conocimiento

Este tipo de función también se conoce como función de energía *estadística*, incluyéndose las funciones entrenadas por aprendizaje de máquina. Utiliza las bases de datos de proteínas conocidas para extraer *pseudo-potenciales* que modelen las interacciones en la proteína. Pueden combinarse con funciones de energía basadas en la física buscando mejorar los resultados logrados por ellas. De hecho, para los problemas de predicción de estructuras, diseño de proteínas y empacamiento de la cadena lateral; la mayoría de los métodos del estado del arte utilizan una combinación de componentes de energía basadas en la física y en conocimiento (Khoury *et al.*, 2014; Li *et al.*, 2013; Colbes *et al.*, 2016). Un ejemplo reciente y detallado es el trabajo de Alford *et al.* (2017), referente a los términos empleados en la función de score de ROSETTA. La cantidad de propuestas es mayor a la del tipo anterior, siendo algunos ejemplos: OPUS-PSP (Lu *et al.*, 2008b), RosettaDesign (Liu y Kuhlman, 2006), DOPE (Shen y SaliProtein, 2006), HPMF (Lin *et al.*, 2007) y SIDEpro (Nagata *et al.*, 2012). Este tipo de funciones puede volver a dividirse en: (i) potenciales atómicos, y (ii) potenciales (semi) residuales (de *grano grueso*) (Lu *et al.*, 2008b).

2.3.2. Plegamiento de proteínas

En las células vivas, las proteínas se ensamblan rápidamente. Por ejemplo, las células de la *Escherichia coli*[22] pueden crear una molécula de proteína biológicamente activa de 100 residuos en aproximadamente 5 segundos (Nelson y Cox, 2004). Entonces surge la pregunta: ¿Cómo llega una cadena polipeptídica a su conformación nativa? La paradoja de Levinthal (1969), mencionada en la Sección 1.3, es un ejercicio mental que indica que las proteínas no se pliegan probando cada conformación posible; sino que siguen un camino definido (al menos parcialmente) que consiste en intermedios entre la proteína desnaturalizada y la estructura nativa (Berg *et al.*, 2012).

El camino de plegamiento de una cadena polipeptídica es complicado y aún no se han determinado los principios que guían este proceso, aunque se han propuesto varios modelos. En uno de ellos, el proceso es jerárquico: se forman primero las estructuras secundarias locales, seguido de interacciones de largo alcance entre distintas

[22] O simplemente *E. coli*: nombre de un tipo de bacteria que normalmente se encuentra en el intestino de los animales de sangre caliente.

regiones para formar así una estructura estable. En un modelo alternativo, el plegamiento inicia con un colapso espontáneo que resulta en una conformación compacta mediante las interacciones hidrofóbicas entre residuos no polares. Esta conformación inicial, conocida como *glóbulo fundido*, puede tener un alto contenido de estructuras secundarias, pero las cadenas laterales de varios residuos no se encuentran fijas.

La mayoría de las proteínas probablemente se pliegan mediante un proceso que incorpora características de ambos modelos. En lugar de seguir un solo camino desde la estructura desnaturalizada a la nativa, un conjunto de proteínas puede tomar una variedad de caminos con el mismo destino, donde el número de diferentes conformaciones parcialmente plegadas decrece a medida que se aproximan a la conformación nativa (Nelson y Cox, 2004). Es por ello que, termodinámicamente, el proceso de plegado puede verse como una especie de *embudo de energía*. Las conformaciones desnaturalizadas se caracterizan por tener una energía relativamente alta. A medida que ocurre el plegamiento, el estrechamiento del embudo representa un decrecimiento en el número de conformaciones con menores valores de energía. En una situación ideal, el embudo no presentaría los mínimos locales a los costados, los cuales representan estados intermedios semi-estables que pueden retrasar el proceso de plegado. Al fondo del embudo, el conjunto de plegamientos intermedios se reduce una sola conformación nativa (o a un conjunto pequeño de conformaciones nativas).

Si se considera el problema de predicción de estructura como un problema de optimización (Boyd y Vandenberghe, 2004), la función de score es la función objetivo que se desea minimizar. Desde el punto de vista computacional, la presencia de los estados intermedios semi-estables (ver Figura 14) indica que la función objetivo es multimodal (ésto es, tiene múltiples óptimos locales), lo que dificulta el proceso de búsqueda (Molga y Smutnicki, 2005).

2.3.3. Enfoques empleados para la determinación *in silico* de estructuras

La determinación computacional de la estructura de una proteína a partir de su secuencia es uno de los problemas no resueltos más importantes en la naturaleza y recientemente ha sobrepasado la barrera de los 50 años (Khoury *et al.*, 2014), aunque desde ese tiempo al presente se han tenido importantes avances (Moult *et al.*, 2016).

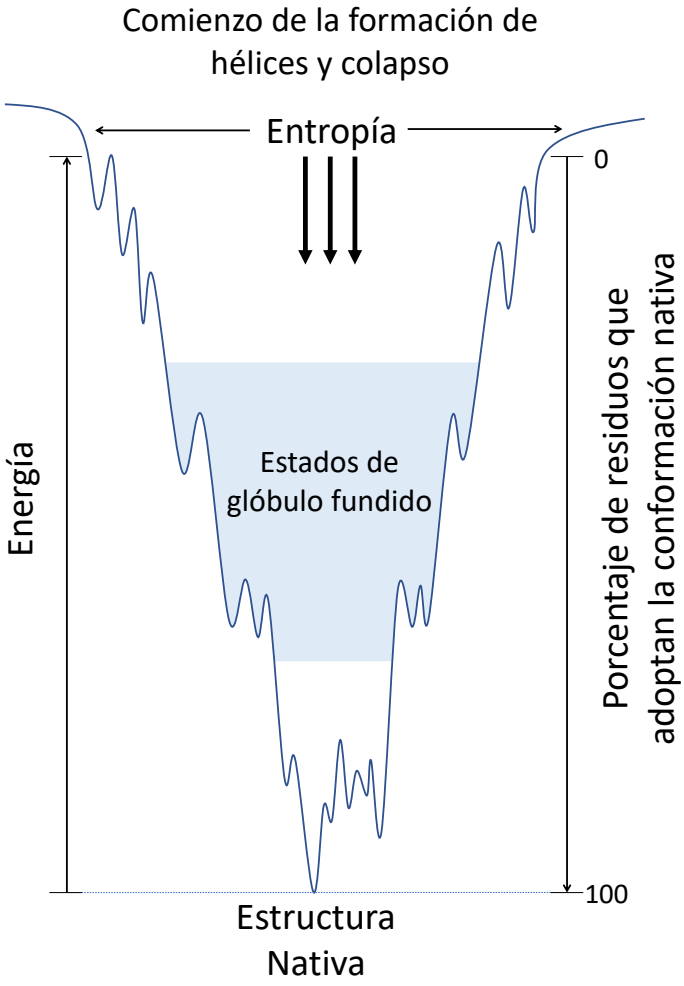

Figura 14. El embudo de energía que ilustra la termodinámica en el plegamiento de proteínas. Imagen adaptada de Nelson y Cox (2004).

El CASP[23] (Moult *et al.*, 1995) es una competencia bianual para evaluar los métodos del estado del arte, y en el año 2016 se realizó la undécima edición. El CASP utiliza *pruebas ciegas* para evaluar el desempeño de los métodos: los participantes reciben secuencias de aminoácidos de estructuras desconocidas y se les pide que depositen modelos estructurales. Estos modelos son comparados con las estructuras determinadas experimentalmente durante el tiempo otorgado para entregar las respuestas.

Existen dos caminos principales para llevar a cabo la predicción *in silico* de estructuras. El primero se basa en la similitud detectable que abarque la mayor parte de la secuencia modelada y al menos una estructura conocida. Aquí entran el *modelado homólogo*, el *enhebrado* y el *modelado comparativo*. El segundo camino, denominado

[23]Critical Assessment of Techniques for Protein Structure Prediction.

de novo[24] o *ab initio*[25], predice la estructura únicamente a partir de la secuencia, sin emplear la similitud de la secuencia modelada con otras de estructura conocida (Baker y Sali, 2001). Ambos métodos usan algoritmos de búsqueda avanzados para alcanzar el óptimo global (o resultados cercanos al mismo). Varias propuestas realizan luego un proceso de *refinamiento* para incrementar la precisión de la predicción (Khoury *et al.*, 2014).

El modelado de una secuencia basado en estructuras conocidas consta de los siguientes cuatro pasos: (i) encontrar estructuras conocidas relacionadas a la secuencia a ser modelada (se las llaman *plantillas*), (ii) alinear la secuencia con las plantillas, (iii) construir el modelo, y (iv) evaluar el modelo (Baker y Sali, 2001). Por lo tanto, el modelado comparativo tiene una fuerte dependencia de estructuras determinadas experimentalmente.

La predicción *ab initio* no tiene tal limitación y por ello se la conoce como “El santo grial” de la predicción de estructuras (Khoury *et al.*, 2014), ya que se basa directamente en la hipótesis termodinámica de Anfinsen. Por lo tanto, se hace una búsqueda a gran escala dentro del espacio de conformaciones posibles de estructuras terciarias que tengan un valor particularmente bajo de energía (o *score*), dada la secuencia de aminoácidos de la proteína que se quiere modelar. Dos factores claves para estos métodos son: (i) un algoritmo que realice la búsqueda de conformaciones de manera eficiente, y (ii) la precisión de la función de score empleada para evaluar tales conformaciones. Para permitir una búsqueda rápida y eficiente en el espacio de conformaciones, se suele representar únicamente un subconjunto de átomos de la cadena de la proteína de manera explícita; y la función de score utilizada debe emplear términos que reflejen la influencia de los átomos omitidos (Baker y Sali, 2001).

Considerando estos dos enfoques principales, la evaluación del desempeño de los métodos del CASP se dividen en dos categorías principales (Moult *et al.*, 2016):

- Modelado basado en plantillas (*Template-Based Modeling*, TBM).
- Modelado libre (*Free Modeling*, FM)

[24] “Desde nuevo”.
[25] “Desde el inicio”.

Los métodos en TBM generalmente tienen un mejor desempeño que los de FM. En el CASP10 (2014), el promedio de los mejores cinco métodos en la categoría FM fue aproximadamente la mitad de la precisión alcanzada por los métodos en TBM (Khoury *et al.*, 2014). La dificultad de cada proteína objetivo en el CASP está dada por una escala en función a la similitud de la secuencia y estructura de dicha proteína con las disponibles en el PDB (Dill y MacCallum, 2012). Las proteínas difíciles de predecir tienen una secuencia muy distinta o directamente una estructura con forma distinta a las almacenadas en el PDB. Para secuencias de entrada con un porcentaje de identidad mayor al 30 % a alguna plantilla, se puede esperar que la estructura predicha tenga una estimación razonable a la topología. Por debajo de este porcentaje, la precisión de la predicción es un problema desafiante (Khoury *et al.*, 2014).

Los métodos FM suelen ser efectivos para proteínas pequeñas, y luego de décadas sin mejoras significativas en la categoría FM, han dado un salto significativo mediante los nuevos métodos de predicción de contactos[26]. Aún así, en la actualidad no existe método que pueda predecir estructuras consistentemente sin el uso de plantillas, posiblemente porque los métodos de búsqueda de conformaciones y las funciones de score son imperfectas (Khoury *et al.*, 2014).

2.4. Ángulos de torsión en la estructura tridimensional

Existen dos formas de representación matemática de la estructura tridimensional de una proteína. Una de ellas es mediante las coordenadas cartesianas, que es la forma empleada para presentar los datos atómicos en los archivos del PDB. Las coordenadas cartesianas son sensibles a la rotación y traslación de la estructura.

La segunda forma es mediante tres medidas entre grupos de átomos unidos por enlaces químicos, éstas son llamadas coordenadas internas: (i) la longitud del enlace (dos átomos), (ii) el ángulo entre enlaces (tres átomos), y el (iii) ángulo de torsión (cuatro átomos). El **ángulo de torsión** en una cadena de átomos A-B-C-D es el ángulo diedro entre el plano que contiene a los átomos A, B y C; y el plano que contiene a B, C y D; tal como se muestra en la Figura 15. A este ángulo se lo conoce también como

[26] Consiste en predecir, a partir de la secuencia, qué residuos están a una distancia umbral dentro de la estructura tridimensional. Las predicciones pueden emplearse como restricciones en el problema de predicción de estructuras, disminuyendo así el espacio de posibles conformaciones.

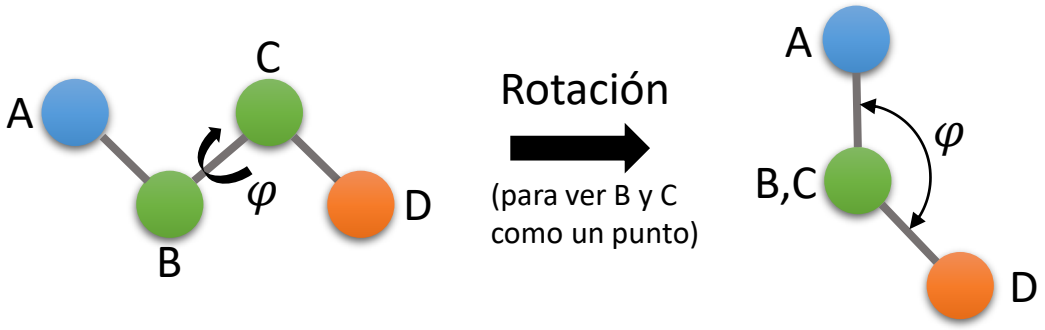

Figura 15. El ángulo de torsión φ en la cadena de átomos A-B-C-D es el ángulo diedro entre los planos determinados por los átomos A-B-C y B-C-D. Se puede ver también como el ángulo de giro del enlace B-C, manteniendo fijas las posiciones de los átomos A, B y C.

el ángulo de rotación del enlace B-C.

Esta representación es invariante a las operaciones de rotación y traslación, aunque el motivo principal de su utilidad no se debe a ello, sino que la naturaleza de los enlaces covalentes en las cadenas polipeptídicas restringe el valor de las longitudes de enlace y ángulos entre enlaces (Berg *et al.*, 2012; Nelson y Cox, 2004). Por lo tanto, los valores de los ángulos de torsión son los que principalmente definen las coordenadas atómicas de una estructura candidata para una proteína. Existen ángulos de torsión en la cadena principal y también en las cadenas laterales de los residuos. Cada residuo cuenta con tres ángulos de torsión en la cadena principal, mientras que el número de ángulos de torsión en su cadena lateral depende del tipo de residuo.

Esta representación también se emplea en los términos relacionados con la descripción química de enlaces covalentes y enlaces de hidrógeno en funciones de energía basadas en la física: como ejemplos se puede citar a AMBER (Cornell *et al.*, 1995) y CHARMM (Brooks *et al.*, 2009).

Pasar de coordenadas cartesianas a longitudes y ángulos es relativamente sencillo; y para el camino opuesto existen varios métodos que se diferencian en la cantidad de operaciones matemáticas necesarias (lo cual afecta el tiempo de ejecución del algoritmo de conversión). El método empleado en el presente trabajo es el de Parsons *et al.* (2005). Solamente se necesitan las posiciones de los primeros tres átomos pesados de la cadena principal (N, C_α y C) para determinar las coordenadas tridimensionales de los demás átomos de la proteína o de la columna vertebral de la proteína.

En relación a los ángulos de torsión en la cadena principal, una característica importante de la cadena polipeptídica es que el enlace peptídico (*C*-*N*) es *plano* (Berg *et al.*, 2012), lo cual limita el número de conformaciones posibles para esta cadena. Así, para un par de residuos unidos por un enlace peptídico, cinco átomos pesados están en un mismo plano: el carbono alfa y el grupo CO del primer residuo, y los átomos N y carbono alfa del segundo residuo. Debido a las características químicas que involucra el enlace peptídico, el mismo no puede rotar libremente. Por lo tanto, la cadena principal de una cadena polipeptídica se puede representar por una serie de planos consecutivos que comparten un punto en común (el carbono alfa), como se observa en la Figura 16 (Nelson y Cox, 2004).

Sin embargo, se permite la rotación alrededor de los enlaces N-C_α (ángulo ϕ) y C_α-C (ángulo ψ); y esta libertad de rotación permite a las proteínas adoptar múltiples conformaciones (o plegamientos). Para un residuo i de la proteína, los ángulos ϕ^i y ψ^i se calculan de la siguiente forma:

- ϕ^i: con la cadena $C^{i-1} - N^i - C^i_\alpha - C^i$
- ψ^i: con la cadena $N^i - C^i_\alpha - C^i - N^{i+1}$

Por convención, estos dos ángulos tienen un valor de 180° cuando la cadena está en una conformación completamente extendida y todos los átomos están en el mismo plano. En principio ϕ y ψ pueden tener cualquier valor entre -180° y +180°, pero ciertos valores están prohibidos debido a *colisiones* entre átomos (Nelson y Cox, 2004). En la Subsección 3.6.4 se discute con mayor detalle el concepto de colisión atómica.

2.4.1. Ángulos de torsión de la cadena lateral

Los ángulos de torsión de la cadena lateral se denotan mediante χ_i, donde i indica la posición de dicho ángulo dentro de la cadena lateral del residuo. La cantidad de ángulos de torsión de la cadena lateral de un residuo depende de su tipo: la Figura 6 muestra los ángulos de torsión para los enlaces que tienen libertad de rotación en cada tipo de residuo, mientras que la Tabla 1 muestra los átomos involucrados en cada ángulo de torsión de cada tipo de residuo. La alanina (ALA) la glicina (GLY) no tienen ángulos de torsión en la cadena lateral.

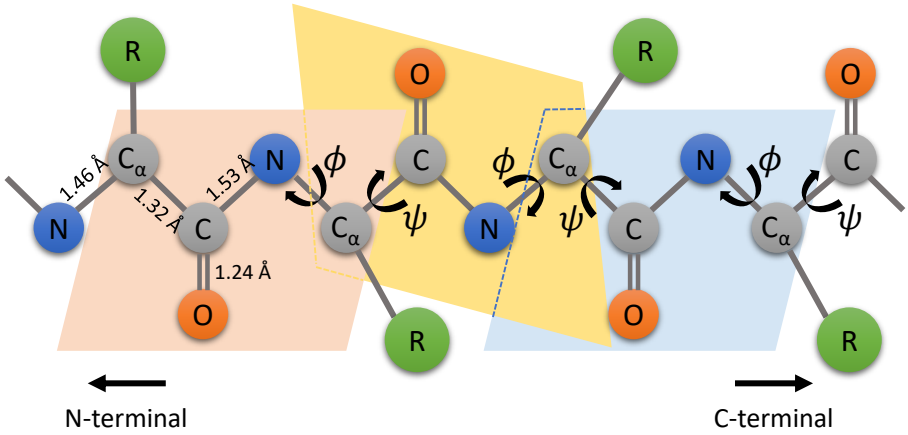

Figura 16. Los enlaces peptídicos no pueden rotar, por lo que los átomos que se encuentran dentro de cada paralelogramo están dentro del mismo plano. Los ángulos ϕ y ψ de cada residuo definen principalmente la conformación de la cadena principal. Además, se muestran las longitudes típicas de los enlaces entre átomos de la cadena principal, de acuerdo a Nelson y Cox (2004).

Los átomos *principales* de la cadena lateral son los átomos de ella necesarios para calcular los ángulos de torsión de la cadena lateral. Si se revisa la composición de cada tipo de residuo en la Figura 6, puede notarse que existen casos en los que no todos los átomos son necesarios para el cálculo de los ángulos de torsión de un residuo en particular. A estos átomos se los conoce como átomos *redundantes* de la cadena lateral. Entonces, si se conoce la ubicación de los átomos principales de la cadena lateral, se puede obtener las coordenadas de los átomos redundantes en ella. Esto se debe a que las propiedades químicas de los enlaces dentro de la cadena lateral restringen las conformaciones que la misma puede adoptar. El procedimiento para obtener las coordenadas de todos los átomos de la cadena lateral a partir de los ángulos de torsión de la cadena lateral, y los parámetros de longitudes de enlaces y ángulos entre enlaces, se detalla en el Anexo A.1.

De la misma forma en que los ángulos de torsión de la cadena principal ϕ y ψ definen la estructura de la proteína a nivel macro, los ángulos de torsión de la cadena lateral de un residuo definen el *empacamiento* del mismo. Así, los ángulos de torsión de todos los residuos definen el **empacamiento de la cadena lateral de la proteína**. El problema de interés en el presente trabajo de investigación es determinar los ángulos de torsión de la cadena lateral de una proteína dada su secuencia de aminoácidos y los átomos de la cadena principal; y este problema se discute con detalle en el Capítulo 3.

Tabla 1. Átomos empleados para calcular los ángulos de torsión para cada tipo de residuo.

Ángulo de torsión	Cadena de átomos	Tipos de residuo
χ_1	$N-C_\alpha-C_\beta-C_\gamma$	ARG,ASN,ASP,GLN,GLU,HIS,ILE,LEU LYS,MET,PHE,PRO,TRP,TYR,VAL
	$N-C_\alpha-C_\beta-O_\gamma$	SER,THR
	$N-C_\alpha-C_\beta-S_\gamma$	CYS
χ_2	$C_\alpha-C_\beta-C_\gamma-C_\delta$	ARG,GLN,GLU,ILE,LEU LYS,PHE,PRO,TRP,TYR
	$C_\alpha-C_\beta-C_\gamma-N_\delta$	HIS
	$C_\alpha-C_\beta-C_\gamma-O_\delta$	ASN,ASP
	$C_\alpha-C_\beta-C_\gamma-S_\delta$	MET
χ_3	$C_\beta-C_\gamma-C_\delta-C_\epsilon$	LYS
	$C_\beta-C_\gamma-C_\delta-N_\epsilon$	ARG
	$C_\beta-C_\gamma-C_\delta-O_\epsilon$	GLN,GLU
	$C_\beta-C_\gamma-S_\delta-C_\epsilon$	MET
χ_4	$C_\gamma-C_\delta-C_\epsilon-N_\zeta$	LYS
	$C_\gamma-C_\delta-N_\epsilon-C_\zeta$	ARG
χ_5	$C_\delta-N_\epsilon-C_\zeta-N_\eta$	ARG

Capítulo 3. Problema de empacamiento de la cadena lateral de proteínas

Este capítulo describe el problema central de este trabajo de investigación: el *problema de empacamiento de la cadena lateral de proteínas*. Se revisan los componentes principales de los métodos que abordan este problema, así como el desempeño de los métodos del estado del arte. Finalmente, se explica la necesidad de contar con métodos de evaluación de funciones de score.

3.1. Definición del PSCPP

El **problema de empacamiento de la cadena lateral** (PSCPP por sus siglas en inglés) consiste en predecir las coordenadas tridimensionales de todos los átomos de las cadenas laterales de los residuos de la proteína, dada su secuencia de aminoácidos y las coordenadas de los átomos de su cadena principal. El PSCPP es importante para varios problemas, entre los que destacan:

- Predicción de estructura de proteínas: el PSCPP es especialmente importante durante el proceso del modelado homólogo en residuos de regiones con baja similitud de secuencia (Blundell *et al.*, 1987; Krieger *et al.*, 2005).
- Diseño de proteínas: para cada secuencia candidata se debe resolver el PSCPP para conocer su valor de energía de acuerdo a la función empleada (Huang *et al.*, 2016). Así, el problema de diseño de proteínas es finalmente una extensión del PSCPP.
- Refinamiento de estructuras cristalográficas: de manera parecida al modelado homólogo, se puede emplear el conocimiento almacenado en el PDB para modelar regiones que no alcanzan una resolución atómica mínima (Jones y Thirup, 1986).

Por ello, los métodos para el PSCPP deben ser tanto precisos como rápidos. Por ejemplo, la mayoría de los métodos de búsqueda para el PSCPP usan un enfoque

Monte Carlo[1] para proponer cambios en la secuencia actual; por lo tanto, el PSCPP se debe resolver para determinar si cada cambio propuesto es favorable o no (Voigt *et al.*, 2000; Liu y Kuhlman, 2006; Krivov *et al.*, 2009; Khoury *et al.*, 2014).

Si bien el PSCPP consiste en determinar las coordenadas cartesianas de todos los átomos de la cadena lateral, las mismas vienen determinadas principalmente por los valores de los ángulos de torsión de la cadena lateral[2]. Como las longitudes y ángulos entre enlaces no sufren una gran variación, se emplean parámetros obtenidos mediante el estudio de estructuras almacenadas en el PDB. Los parámetros utilizados en este trabajo son los de Engh y Huber (1991).

Entonces, el PSCPP se reduce a encontrar los valores de los ángulos de torsión de la cadena lateral de cada residuo. Como los ángulos son valores continuos, en principio existe un número infinito de conformaciones en el espacio de búsqueda. Un primer intento de discretización de este espacio podría ser la consideración de intervalos regulares en los valores de los ángulos de torsión. Sin embargo, varios trabajos que se centraron en el análisis estadístico de las conformaciones de las cadenas laterales de estructuras conocidas encontraron que, para cada uno de los 20 aminoácidos naturales, los valores de los ángulos de torsión no están uniformemente distribuidos (Chandrasekaran y Ramachandran, 1970; Richards, 1977; Bhat *et al.*, 1978; Janin y Wodak, 1978; James y Sielecki, 1983). Estos trabajos muestran que existen ciertas regiones en el espacio de búsqueda en donde se agrupan las conformaciones de las cadenas laterales de las proteínas consideradas. Ésto dio lugar a la discretización del espacio de búsqueda mediante la aparición de las llamadas *bibliotecas de rotámeros* (Dunbrack, 2002).

Debido a la importancia del PSCPP, existe una gran cantidad de métodos propuestos para el mismo durante las dos últimas décadas (Canutescu *et al.*, 2003; Lu *et al.*, 2008a; Krivov *et al.*, 2009; Cao *et al.*, 2011; Miao *et al.*, 2011; Nagata *et al.*, 2012; Francis-Lyon y Koehl, 2014; Liang *et al.*, 2011a; Quan *et al.*, 2014; Ryu *et al.*, 2016; Gaillard *et al.*, 2016). La mayoría de estos trabajos formulan el PSCPP como un problema de optimización combinatoria, donde el espacio de búsqueda depende del conjun-

[1] En este método, se realizan cambios sucesivos al azar de una solución en particular (que para los problemas considerados aquí puede ser una secuencia o estructura). Si el cambio es favorable en términos de la función objetivo (la energía en este caso), el movimiento se acepta. En caso contrario, se usa la probabilidad de Boltzmann para determinar si el cambio se acepta o no.

[2] Ver la Sección 2.4

to discreto de posibles conformaciones para cada residuo, conocidos como *rotámeros*. Siguiendo este enfoque, los componentes principales para este problema son: (i) una biblioteca de rotámeros, (ii) una función de score, y (iii) un algoritmo de búsqueda. De esta forma, el PSCPP se convierte en un problema de optimización combinatoria que consiste en encontrar un conjunto de rotámeros de la biblioteca (un rotámero para cada residuo de la proteína) que minimice una función de score dada. Akutsu (1997) demostró que este problema es NP-difícil, por lo que no se conoce un algoritmo eficiente para encontrar la solución óptima del PSCPP.

3.2. Biblioteca de rotámeros

Un **rotámero** es una conformación de la cadena lateral representada por un conjunto de valores, uno para cada grado de libertad de rotación (ángulo de torsión). El término rotámero proviene del inglés *rotamer*, que a su vez es una abreviatura de ***rot****ational iso****mer*** (isómero rotacional). La biblioteca de rotámeros es una colección de rotámeros para cada tipo de residuo (Dunbrack, 2002). Las bibliotecas de rotámeros usualmente contienen información acerca de la conformación y la frecuencia (o *probabilidad*) de dicha conformación. También pueden poseer información sobre la varianza en relación a la media o la moda de los ángulos de torsión para cada uno de los rotámeros.

Las bibliotecas de rotámeros usualmente derivan de análisis estadísticos de conformaciones de las cadenas laterales en estructuras conocidas de proteínas. Dos enfoques empleados son: (i) por agrupamiento de conformaciones (*clustering*); o (ii) mediante la división del espacio de conformaciones en regiones, determinando la conformación promedio en cada región. Debido a la creciente información disponible en el PDB, las bibliotecas se han vuelto cada vez más precisas (Dunbrack, 2002).

El tamaño de la biblioteca empleada y el número de residuos de la proteína determinan el tamaño del espacio de búsqueda. Una biblioteca muy grande puede dificultar el proceso de búsqueda por el tamaño del espacio asociado; mientras que una biblioteca muy pequeña puede no contener ciertas conformaciones que permitan obtener el mínimo de la función de score empleada.

Considerando la dependencia de los ángulos de torsión y frecuencias de los rotá-

meros en relación a la conformación local de la cadena principal, las bibliotecas de rotámeros se dividen en (Dunbrack, 2002):

- Bibliotecas independientes de la cadena principal: no hacen referencia a la conformación de la cadena principal y se calculan a partir de todas las cadenas laterales disponibles para un cierto tipo de residuo. Las primeras bibliotecas de rotámeros publicadas fueron de este tipo (Chandrasekaran y Ramachandran, 1970; Bhat *et al.*, 1978; Ponder y Richards, 1987).
- Bibliotecas dependientes de la estructura secundaria: Janin y Wodak (1978) mostraron la dependencia de la conformación de las cadenas laterales de un residuo en función a la estructura secundaria, proponiendo la primera biblioteca de este tipo. Por lo tanto, estas bibliotecas contienen rotámeros con distintos valores de ángulos de torsión y/o frecuencias dependiendo de la pertenencia del residuo a alguna estructura secundaria (hélices α, láminas β, etc.).
- Bibliotecas dependientes de la cadena principal: Dunbrack y Karplus (1993, 1994) demostraron que los residuos tienen preferencias en su conformación de la cadena lateral en función a los ángulos de torsión de la cadena principal ϕ y ψ, proponiendo así la primera biblioteca de este tipo. En este caso, existe un conjunto de rotámeros de distintos valores de ángulos de torsión y/o frecuencias para cada región en los valores de ϕ y ψ (por ejemplo, cada 10° o 20°). Debido a la precisión que logran las bibliotecas de este tipo, son las más empleadas en los métodos del estado del arte para el PSCPP.

Un ejemplo (parcial) de una biblioteca de rotámeros dependiente de la cadena principal se muestra en la Figura 17.

3.3. Función de Score

Al modelar el PSCPP como un problema de optimización, la función de score es la función objetivo que se quiere minimizar. Como se mencionó anteriormente, las funciones de score modelan las interacciones atómicas que ocurren dentro de la proteína y con el medio en que se encuentra. Existen diferentes tipos de interacciones fisícoquímicas que en conjunto determinan la estructura de una proteína, y la precisión de

Ángulos de torsión del BB

Frecuencia de aparición

Ángulos de torsión del SC

Desviaciones estándar

Residuo	ϕ	ψ	Probabilidad	χ_1	χ_2	χ_3	χ_4	$\sigma(\chi_1)$	$\sigma(\chi_2)$	$\sigma(\chi_3)$	$\sigma(\chi_4)$
ARG	60	-130	0.227859	-63.7	-176.8	-175.5	170.2	8	11.6	11.5	19.8
ARG	60	-130	0.162569	-69.2	-176.2	-177.7	-80.9	7.6	10.8	10.9	9.6
ARG	60	-130	0.08766	-66.6	-179	-66.2	175.2	8.9	13.6	13.1	21.2
ARG	60	-130	0.080352	-66	179.4	65.3	-168.6	11.3	15.5	11.9	20.1
ARG	60	-130	0.070213	-66.9	-170.3	-63.4	-84.2	10	10	11.9	9.8
ARG	60	-130	0.063172	-65.9	-179.9	178.5	87	8.5	12.1	13	13.9
ARG	60	-130	0.051581	-63.9	179.4	67.7	84	9.4	10.2	9.8	10.1
ARG	60	-120	0.140073	-69.5	-176.1	-177.9	-80.9	7.8	11.1	11.2	9.7
ARG	60	-120	0.127271	-65.7	-179.7	177.2	88	8	11.4	12.4	13.2
ARG	60	-120	0.115186	-65.1	-178.2	-176.6	174.5	8.6	11.6	12.4	17.5
ARG	60	-120	0.088579	-171.4	173.5	174.3	80.6	10.5	10.5	13.4	12.3
ARG	60	-120	0.069865	-66.3	-178.1	-66.3	175.6	8.9	14	12.9	21.2
ARG	60	-120	0.06404	-66.1	179.1	65.6	-168.8	11.5	15.7	12.1	20.5
ARG	60	-120	0.055959	-67.1	-171.4	-63.8	-84	10.2	10.3	12.1	10.2
ARG	60	-110	0.135065	-65.9	-179.8	176.6	88.2	7.9	11.3	12.5	13.1
ARG	60	-110	0.118352	-66.5	179.6	-177.7	176	9.3	12.3	13.8	19.3
ARG	60	-110	0.098746	-171.6	173.6	174.3	80.7	10.5	10.6	13.6	12.4
ARG	60	-110	0.071785	-66.5	-177.8	-66.8	175.2	8.8	14.3	12.6	21
ARG	60	-110	0.065801	-66.3	178.8	65.5	-169	11.7	16	12.4	21
ARG	60	-110	0.064886	-68.6	-178.4	-176.8	-81.8	9.1	12.3	13	11.4
ARG	60	-110	0.057497	-67.8	-171.5	-64.3	-84.7	10.1	10.7	12.3	10.2

Figura 17. Ejemplo de la presentación de una biblioteca de rotámeros dependiente de la columna vertebral. BB (*backbone*): cadena principal. SC (*side-chain*): cadena lateral. Los colores representan la agrupación de los rotámeros de acuerdo a los valores de ϕ y ψ.

una función de score está dada por el número de interacciones que considera y el nivel de descripción de las mismas. Por ejemplo:

- Algunas funciones de score consideran cada átomo de cada residuo; mientras que otras consideran todos los átomos como un solo elemento representativo del residuo.
- Otras funciones no consideran las interacciones de los átomos de la proteína con el medio en que se encuentra (por lo general, el agua).

Cabe resaltar que generalmente las funciones de score para el PSCPP[3] son únicamente modelos *aproximados* o *simplificados* para calcular la energía, y que un mayor nivel de precisión en las interacciones consideradas usualmente conlleva una mayor

[3]También se da lo mismo en la predicción de estructuras y diseño de proteínas.

complejidad de la función de score, lo cual finalmente requiere un mayor tiempo de cómputo.

Dirac (1929) ya estableció esta relación entre precisión y tiempo de cómputo casi un siglo atrás, refiriéndose a la mecánica cuántica: "*Las leyes físicas subyacentes necesarias para la teoría matemática de una gran parte de la física y la totalidad de la química son* ***completamente conocidas****, y la dificultad es únicamente que la aplicación exacta de estas leyes conduce a ecuaciones* ***demasiado complicadas para ser solubles****. Por lo tanto, resulta deseable que se desarrollen* ***métodos prácticos aproximados*** *de aplicación de la mecánica cuántica, lo que puede conducir a una explicación de las* ***principales características*** *de los sistemas atómicos complejos* ***sin demasiados cálculos****.*"

3.3.1. ¿Dinámica molecular para el PSCPP?

Resulta oportuno considerar la posibilidad de emplear simulaciones de dinámica molecular para el PSCPP, o bien para el problema de predicción de estructura de proteínas; ya que las funciones de energía empleadas en la dinámica molecular tienen una muy alta precisión. La *dinámica molecular* es una simulación computacional para estudiar el movimiento de un conjunto de átomos, que resulta ser una aplicación de un problema conocido como el *problema de los N cuerpos* (Rapaport, 2004). Este problema se originó en la dinámica del sistema solar, y resolverlo de forma exacta es intratable para el caso de tres o más cuerpos.

Por lo tanto, para la dinámica molecular se emplean aproximaciones mediante métodos numéricos, calculándose paso a paso las soluciones de las ecuaciones clásicas de movimiento (Allen *et al.*, 2004). Así, en un modelo simple se considera, en cada intervalo Δt, lo siguiente para cada átomo i del sistema de N átomos[4]:

$$m_i \boldsymbol{\alpha}_i = \mathbf{f}_i \qquad \mathbf{f}_i = -\frac{\partial \mathcal{U}(\mathbf{r}^N)}{\partial \mathbf{r}_i} \tag{2}$$

Donde m_i es la masa del átomo i, $\boldsymbol{\alpha}_i$ es su aceleración debido a la fuerza $\mathbf{f}_i$; la cual a su vez depende de su posición $\mathbf{r}_i$ y la función de energía $\mathcal{U}(\mathbf{r}^N)$ que considera la

[4] Atómos de la proteína y el solvente.

interacción con los demás átomos del sistema (donde $\mathbf{r}^N = (\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N)$). Por ello, para cada Δt se necesita calcular la fuerza f_i que actúa sobre cada átomo i, actualizando sus valores de posición, velocidad y aceleración. El resultado de una simulación de dinámica molecular son las posiciones y velocidades de cada uno de los átomos del sistema considerado, para cada intervalo Δt durante la simulación (Allen *et al.*, 2004). La precisión de los resultados de la dinámica molecular depende de muchos factores, entre los cuales destacan la precisión de las funciones de energía, el valor de Δt (usualmente en el orden de los femtosegundos) y los correctores empleados para considerar la discretización del tiempo. El tiempo de simulación necesario para producir resultados confiables depende del tiempo requerido para lograr la *convergencia* o equilibrio del sistema analizado (Allen *et al.*, 2004), y así utilizar los valores promedio de posición en este estado.

La dinámica molecular puede considerarse el método más preciso para simular las interacciones que ocurren entre los átomos del sistema proteína+solvente, por lo que surge la pregunta: ¿Por qué no aplicarla para el PSCPP o la predicción de estructuras? La razón principal está dada por el tiempo necesario para realizar las simulaciones: para proteínas pequeñas (<100 residuos), actualmente se alcanzan tiempos de simulación en el orden de los milisegundos (Vendruscolo y Dobson, 2011). El plegamiento de las proteínas se da generalmente en el orden de los segundos (Berg *et al.*, 2012), pero para ciertas proteínas con un plegamiento rápido (en el orden de los milisegundos) se ha podido simular el proceso de plegamiento mediante dinámica molecular (Khoury *et al.*, 2014). Para proteínas más grandes generalmente se logra simular en el orden de los nanosegundos, por lo que la dinámica molecular se emplea principalmente para el refinamiento de las estructuras predichas *in silico* y el análisis de su estabilidad (Moult *et al.*, 2016).

3.3.2. Términos frecuentes en las funciones de score para el PSCPP

Como se mencionó anteriormente, los métodos para el PSCPP deben ser rápidos; y sumado al hecho de que ya se conoce la estructura de la cadena principal, las funciones de score empleadas son más simples que en el caso de la predicción de estructuras. Para implementación computacional más rápida, las funciones de score generalmente se limitan a términos que consideran interacciones internas en un resi-

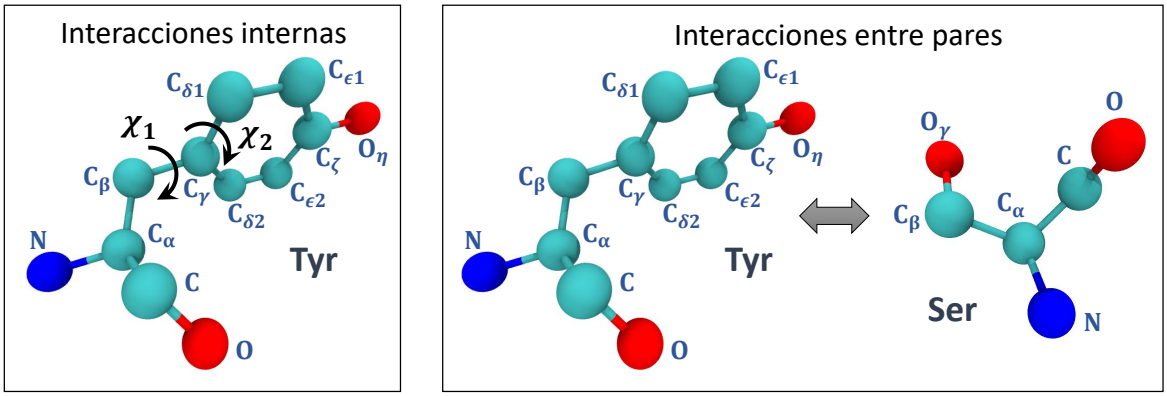

Figura 18. Tipos de interacciones consideradas en el PSCPP. Las interacciones internas generalmente dependen de los ángulos de torsión de la cadena lateral, o están relacionadas a las probabilidades de los rotámeros en la biblioteca. Las interacciones entre pares de residuos consideran las interacciones de Van der Waals, enlaces de hidrógeno, enlaces disulfuro y otros. Imagen obtenida mediante el software VMD (Humphrey *et al.*, 1996).

duo o interacciones entre pares de residuos, como se muestra en la Figura 18. Estas interacciones generalmente se pre-calculan y almacenan en una estructura de datos, lo cual permite una evaluación rápida del score total durante el proceso de búsqueda de soluciones. Los términos que generalmente se emplean dentro de las funciones de los métodos del estado del arte se describen a continuación.

3.3.2.1. Interacciones de Van der Waals

Estas interacciones no-covalentes[5] aparecen básicamente debido a que la distribución de la carga electrónica alrededor de un átomo fluctúa con el tiempo. En cualquier instante, la distribución de carga no es perfectamente simétrica; lo cual provoca interacciones electrostáticas con sus átomos vecinos para inducir asimetrías complementarias en la distribución electrónica. El átomo y sus vecinos se atraen entre sí, y esta atracción aumenta a medida que dos átomos se acercan, hasta que los mismos están separados a una distancia denominada *distancia de contacto*. A distancias menores que la de contacto, fuerzas repulsivas se vuelven dominantes debido a la superposición de las nubes de electrones exteriores de ambos átomos.

La Figura 19 es una representación esquemática de las interacciones de Van der Waals para dos átomos. Existen varias funciones matemáticas que buscan modelar estas interacciones entre un par de átomos, siendo el potencial de Lennard-Jones (1924)

[5] Interacciones entre átomos que no están unidos por un enlace covalente.

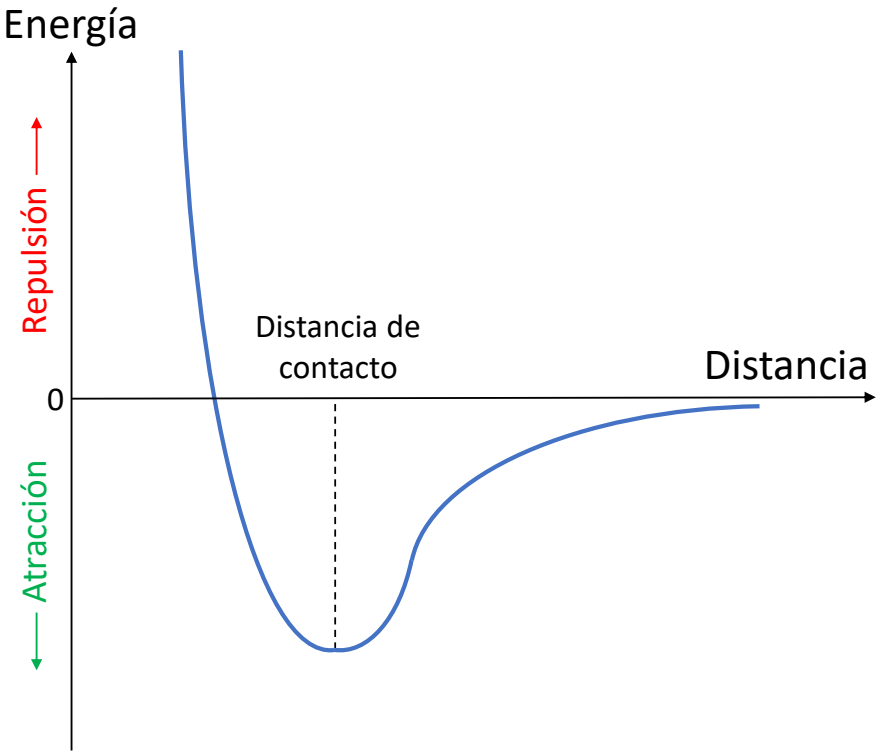

Figura 19. Representación de las interacciones de Van der Waals. La atracción entre los dos átomos es máxima en la distancia de contacto, y a distancias menores empieza a crecer rápidamente la fuerza de repulsión debido a la cercanía de sus nubes de electrones. Adaptación de la imagen en (Berg *et al.*, 2012).

una de las representaciones más empleadas. El mismo tiene la forma:

$$E_{vDW-LJ}(d) = 4\epsilon\left[\left(\frac{\sigma}{d}\right)^{12} - \left(\frac{\sigma}{d}\right)^{6}\right] \quad \text{o bien:} \quad E_{vDW-LJ}(d) = \frac{A}{d^{12}} - \frac{B}{d^{6}} \qquad (3)$$

Donde d es la distancia entre los átomos, ϵ es la *profundidad* determinada por el valor de la energía en la distancia de contacto, y σ es la distancia en donde la energía es nula. La segunda forma es la más empleada en simulaciones computacionales, donde A y B están en función de ϵ y σ.

3.3.2.2. Frecuencia de rotámeros

Este término se basa en el conocimiento y se relaciona directamente con las probabilidades (o frecuencias) de los rotámeros en la biblioteca empleada. Básicamente favorece a las conformaciones de las cadenas laterales que más aparecen en las proteínas con estructura conocida. Una forma común de expresión de este término, para

un determinado residuo i, es:

$$E_{Rot}(r_i) = -\gamma_i \log\left(\frac{p(r_i)}{p(r_{maxp})}\right) \quad (4)$$

Donde γ_i es una constante que depende del tipo del residuo i al que corresponde el rotámero r_i, $p(r_i)$ es la probabilidad del rotámero de la biblioteca; y $p(r_{maxp})$ es la mayor probabilidad dentro de los rotámeros en la biblioteca para el tipo del residuo i. El uso de $p(r_{maxp})$ es opcional. También puede darse el caso de no considerar todos los rotámeros para un cierto tipo de residuo, sino sólo los más probables. En este caso se puede agregar un factor para normalizar los valores de probabilidad de los rotámeros empleados.

3.3.2.3. Enlaces de hidrógeno

Estas interacciones son básicamente de tipo electrostático. Cuando un átomo de hidrógeno se encuentra unido mediante enlace covalente con otro átomo altamente electronegativo (por ejemplo, el oxígeno), no existe una distribución justa de los electrones que comparten ambos átomos; ya que los electrones generalmente se encuentran más cerca del átomo más electronegativo. Como resultado se forma un dipolo, donde cada hidrógeno tiene una carga parcial positiva (δ^+) y el átomo electronegativo tiene una carga parcial negativa (δ^-). Por lo tanto, el átomo de hidrógeno puede interactuar con otro átomo que tiene una carga parcial negativa (que es otro átomo altamente electronegativo) a través de una interacción electrostática (Berg *et al.*, 2012).

El *donador* en un enlace de hidrógeno es el grupo que incluye el átomo de hidrógeno y el átomo electronegativo unido a él mediante enlace covalente, mientras que el otro átomo electronegativo es el *aceptor*. Los átomos electronegativos que participan en estos enlaces son el oxígeno, el nitrógeno y el flúor.

Cada enlace peptídico de una cadena polipeptídica tiene un donador (grupo NH) y un aceptor (grupo CO) para un enlace de hidrógeno, y los enlaces de hidrógeno entre estos grupos de la cadena principal son una característica distintiva de su estructura. Los enlaces de hidrógeno son principalmente responsables de la formación de los elementos de la estructura secundaria. Los residuos de aminoácidos polares contienen

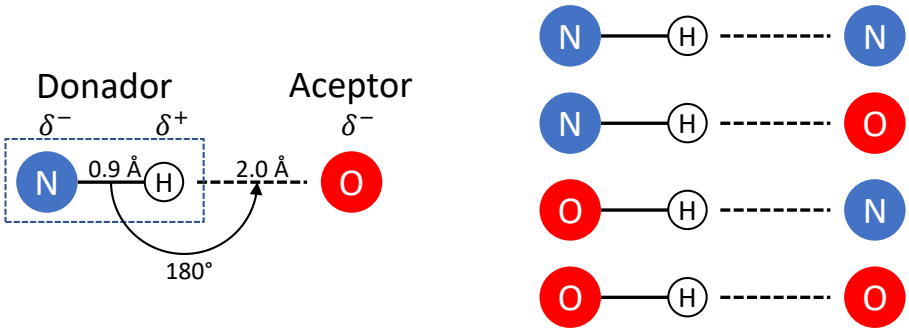

Figura 20. Enlaces de hidrógeno en proteínas. Los enlaces de hidrógeno se indican mediante las líneas discontinuas y tienen menor fuerza que los enlaces covalentes. Adaptación de la imagen en (Berg *et al.*, 2012).

grupos funcionales que forman enlaces de hidrógeno con el agua, mientras que los grupos NH y CO de los residuos hidrofóbicos que pertenecen a un núcleo hidrofóbico forman enlaces de hidrógeno entre ellos (Berg *et al.*, 2012). La Figura 20 muestra los enlaces de este tipo que pueden formarse en el caso de las proteínas.

Los enlaces de hidrógeno y las interacciones de Van der Waals son individualmente débiles (significativamente más débiles que los enlaces covalentes), pero colectivamente tienen una fuerte influencia en la estructura tridimensional de las proteínas (Nelson y Cox, 2004).

3.3.2.4. Enlaces disulfuro

Los enlaces disulfuro son enlaces covalentes fuertes que se forman por la oxidación de dos residuos de cisteína, formándose así un residuo dímero llamado *cistina* (Ver Figura 21). Los residuos unidos por enlaces disulfuro son altamente hidrofóbicos; y estos enlaces juegan un papel central en las estructuras de muchas proteínas al formar enlaces covalentes entre distintas partes de la cadena polipeptídica (terciaria), o entre dos cadenas polipeptídicas (cuaternaria) (Nelson y Cox, 2004). De hecho, este tipo de enlaces jugaron un papel muy importante en el experimento de Anfinsen (1973) para establecer la hipótesis termodinámica.

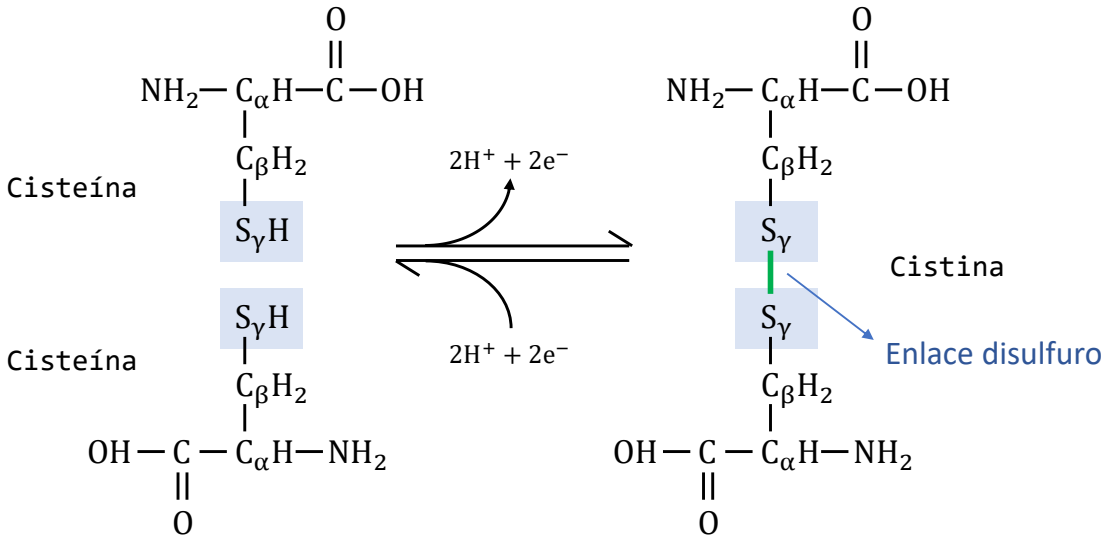

Figura 21. Enlaces disulfuro en proteínas. Los enlaces disulfuro son reversibles y se forman por la oxidación de dos cisteínas. Estos enlaces estabilizan la estructura tridimensional de muchas proteínas. Adaptación de la imagen en (Nelson y Cox, 2004).

3.3.2.5. Consideraciones sobre otros términos

Las funciones más simples, que se limitan a considerar las interacciones de Van der Waals y las preferencias de rotámeros de la biblioteca, proveen resultados relativamente buenos en general y excelentes resultados para el caso de residuos enterrados no polares. El problema es que este enfoque no da resultados precisos para residuos expuestos o parcialmente expuestos al solvente, y tampoco para residuos polares enterrados (Xiang y Honig, 2001; Liang y Grishin, 2002; Lu *et al.*, 2008a).

En la Subsección 3.3.1 se establece la inviabilidad de realizar simulaciones de dinámica molecular para dar soluciones al PSCPP, pero al menos se podría considerar el uso de las funciones de energía que se emplean en ella. Aún así, se ha determinado que el uso de términos que consideren interacciones electrostáticas, enlaces de hidrógeno y otros que aparecen en las funciones de energía empleadas en la dinámica molecular como AMBER (Cornell *et al.*, 1995) y CHARMM (Brooks *et al.*, 2009), no producen mejoras significativas en el PSCPP (Wilson *et al.*, 1993; Petrella *et al.*, 1998; Xiang y Honig, 2001; Liang y Grishin, 2002). Recientemente, Gaillard *et al.* (2016) evaluaron el desempeño de una función de energía a la que llamaron *MMGBSA*, la cual combina términos de AMBER y un modelo para representar al solvente. Si bien esta función se propuso para el diseño de proteínas, se empleó para el PSCPP ya que es

un subproblema que permite una evaluación *in silico*. Los resultados obtenidos están a la par de los métodos del estado del arte para el PSCPP, con la desventaja de que el tiempo de operación necesario es mucho mayor.

De hecho, Liang y Grishin (2002) diseñaron una función de score para el PSCPP que logró mejores resultados que los obtenidos mediante las funciones de energía de AMBER y CHARMM, lo cual indica que deben desarrollarse términos de score orientados únicamente a las conformaciones de la cadena lateral de la proteína. En este sentido, se han desarrollado varios términos basados en el conocimiento. Algunos consideran la orientación de las cadenas laterales (Lu *et al.*, 2008b; Liang *et al.*, 2011a), otros definen funciones basadas en la estadística de contactos atómicos en estructuras conocidas para seleccionar los rotámeros (Samudrala y Moult, 1998); aunque las mejoras que logran no son significativas.

Un punto muy importante es la consideración del solvente. Se ha indicado que la inclusión explícita de las moléculas de agua individuales mejora la precisión para el PSCPP, lo cual se debe básicamente a una representación más realista de los efectos del empacamiento. Por lo tanto, una mejora en el modelado de los efectos del solvente es importante para los residuos expuestos, aunque un modelo muy preciso puede aumentar el costo computacional (Xiang y Honig, 2001). De hecho, el método que tiene el mejor desempeño para el PSCPP tiene un término en su función de score que aproxima la interacción con el solvente; aunque se aplica únicamente en la última fase de la búsqueda debido a su alto costo computacional (Lu *et al.*, 2008a).

3.4. Formulación matemática del PSCPP

El PSCPP, así como el problema de predicción de estructuras, es un problema de interacciones de N cuerpos. Al emplear el enfoque de las bibliotecas de rotámeros, el PSCPP se convierte en un problema de optimización combinatoria, donde se debe elegir una combinación de los rotámeros de la biblioteca (un rotámero para cada residuo) que resulte en una estructura de valor mínimo para una función de score dada. Bajo esta definición, una solución candidata puede representarse como un arreglo de números enteros, donde el valor en una posición i indica la posición en la biblioteca del rotámero asignado al residuo i; como se muestra en la Figura 22.

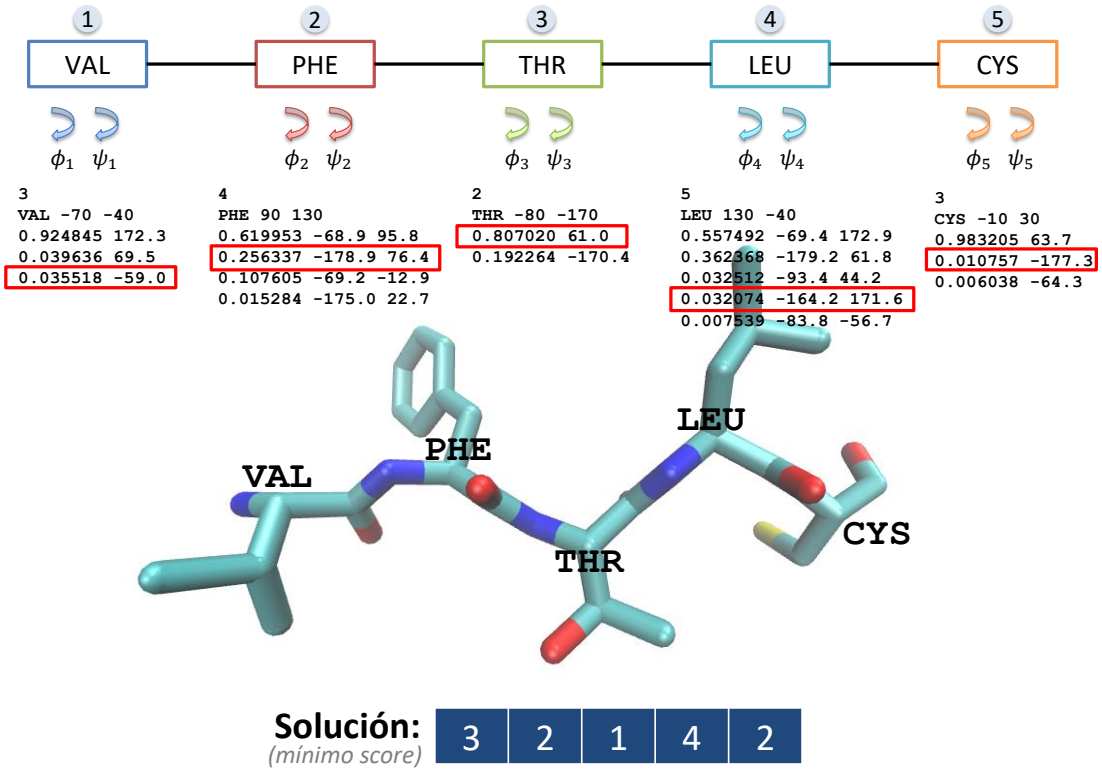

Figura 22. Ejemplo que muestra el PSCPP como un problema combinatorio, considerando una proteína de cinco residuos. Cada residuo tiene un número de opciones determinadas por la biblioteca de rotámeros; y para cada residuo se debe seleccionar un rotámero de tal forma que el conjunto seleccionado sea el de valor mínimo en la función de score empleada.

El PSCPP asociado a una biblioteca de rotámeros independiente de la columna vertebral se define como: dada una secuencia de n aminoácidos $\mathbf{a} = (a_1 a_2 \dots a_n)$, las coordenadas de los átomos de la cadena principal $\vec{c} = (\vec{c}_1, \vec{c}_2, \dots, \vec{c}_n)$, la biblioteca de rotámeros rl y una función de score E; el problema consiste en encontrar los rotámeros $\vec{r}^* = (\vec{r}_1^*, \vec{r}_2^*, \dots, \vec{r}_n^*)$, de tal manera que la función de score $E(\mathbf{a}, \vec{c}, rl, \vec{r})$ sea mínima, donde:

- $a_i \in \Sigma$, siendo Σ el alfabeto de los 20 aminoácidos naturales.
- $\vec{c}_i = \{\vec{N}_i, \vec{C}_i^{\alpha}, \vec{C}_i, \vec{O}_i\}$, siendo $\vec{N}_i, \vec{C}_i^{\alpha}, \vec{C}_i, \vec{O}_i \in \mathbb{R}^3$ las coordenadas tridimensionales de los átomos pesados de la cadena principal (nitrógeno, carbono alfa, carbono carbonilo y oxígeno) del residuo i.
- $\vec{r}_i \in rl(a_i)$ es uno de los rotámeros disponibles para el residuo i.

Si la biblioteca de rotámeros es dependiente de la cadena principal, entonces $rl(a_i)$

cambia a $rl(\alpha_i, \phi_i, \psi_i)$, donde ϕ_i y ψ_i son los ángulos de torsión de la cadena principal correspondientes al residuo i.

Akutsu (1997) demostró que el PSCPP bajo esta formulación es un problema NP-difícil, mediante la reducción del 3SAT (Garey y Johnson, 1979). En el 3SAT, dada una colección $C = \{c_1, \ldots, c_m\}$ de cláusulas sobre un conjunto $V = \{v_1, \ldots, v_n\}$ de variables binarias, se debe decidir si existe o no una asignación para V que satisfaga todas las cláusulas en C, donde cada cláusula consiste de tres literales.

3.5. Algoritmo de búsqueda

Debido a que el PSCPP es un problema NP-difícil (Akutsu, 1997; Xiang y Honig, 2001), se tienen tres posibilidades para abordar problemas de este tipo: (i) usar la fuerza bruta u otro método exacto sin preocuparse por el tiempo de ejecución, (ii) aplicar un algoritmo de aproximación de tiempo polinomial, o (iii) aplicar alguna heurística o metaheurística sin garantía alguna en términos de calidad de solución o tiempo de ejecución (Aarts y Lenstra, 2003). La mayoría de los métodos del estado del arte para el PSCPP utiliza el tercer enfoque, estableciendo una cota máxima para el tiempo de ejecución.

A pesar de reducir el espacio de búsqueda con el uso de las bibliotecas de rotámeros, un par de décadas atrás se creía que la naturaleza de este problema combinatorio era el obstáculo principal para la correcta predicción del empacamiento de la cadena lateral (Lee y Subbiah, 1991; Petrella *et al.*, 1998). Por ello, los trabajos se centraron en proponer métodos de búsqueda para el problema combinatorio (Voigt *et al.*, 2000; Liang y Grishin, 2002). Se emplearon enfoques tanto determinísticos como estocásticos, entre los cuales se pueden citar: algoritmos Monte Carlo y de recocido simulado (Holm y Sander, 1992; Lu *et al.*, 2008a; Krivov *et al.*, 2009; Miao *et al.*, 2011; Liang *et al.*, 2011a), variantes de algoritmos de búsqueda local (Cao *et al.*, 2011), algoritmos de ramificación y acotamiento (Canutescu *et al.*, 2003; Krivov *et al.*, 2009; Miao *et al.*, 2011), redes neuronales (Nagata *et al.*, 2012), el algoritmo de campo medio autoconsistente (SCMF por *self-consistent mean field*) (Koehl y Delarue, 1994; Koehl *et al.*, 2011; Francis-Lyon y Koehl, 2014), algoritmos genéticos (Desjarlais y Handel, 1995; Comte *et al.*, 2011) y algoritmos de colonia de hormigas (Quan *et al.*, 2014).

Existe además un método de *eliminación de callejones sin salida* (DEE por *Dead-end elimination*) que permite reducir el espacio de búsqueda eliminando rotámeros que no pueden pertenecer a la estructura de mínimo score (Desmet *et al.*, 1992).

En el presente trabajo se hace un análisis de varios algoritmos basados en búsqueda local, los cuales se definen con detalle en la Sección 4.3.

3.6. Medidas de calidad en el PSCPP

La evaluación de un método para el PSCPP se realiza *in silico*: se toma una estructura del PDB y se remueven sus átomos de la cadena lateral, luego se realiza la predicción; y finalmente se compara la estructura predicha por el método con la estructura experimental mediante una o varias métricas de calidad. Este procedimiento puede realizarse para un conjunto de proteínas de prueba (o *datasets*), con el fin de evaluar la calidad de la predicción para distintos tipos de estructuras. Las medidas de calidad empleadas en el PSCPP se describen a continuación.

3.6.1. Precisión absoluta

Es la métrica más empleada, y básicamente indica el porcentaje de ángulos de torsión de la cadena lateral predichos correctamente. En la misma se utilizan los valores $\chi_1(\%)$ y $\chi_{1+2}(\%)$, donde $\chi_1(\%)$ representa el porcentaje de residuos cuyos ángulos de torsión χ_1[6] son correctos, mientras que $\chi_{1+2}(\%)$ es el porcentaje de residuos cuyos ángulos de torsión χ_1 y χ_2[7] son ambos correctos. Se considera que un ángulo de torsión es *correcto* si se encuentra alejado a un máximo de 40° del ángulo de torsión de la estructura experimental de la proteína. Este valor arbitrario se ha venido utilizando desde los primeros trabajos que tratan sobre el PSCPP (Summers y Karplus, 1989; Lee y Subbiah, 1991; Wendoloski y Salemme, 1992; Dunbrack y Karplus, 1993), bajo la suposición de que los ángulos en este intervalo corresponden al mismo mínimo de energía. De hecho, los algoritmos con mejores resultados para este problema siguen utilizando este valor como referencia. Los átomos más allá de C_δ tienen relativamente

[6] Ángulo de torsión entre el plano determinado por N, C_α, C_β y el determinado por $C_\alpha, C_\beta, X_\gamma$, donde X_γ depende del tipo de residuo considerado (ver Tabla 1).

[7] Ángulo de torsión entre el plano determinado por $C_\alpha, C_\beta, X_\gamma$ y el determinado por $C_\beta, X_\gamma, X_\delta$, donde X_γ y X_δ dependen del tipo de residuo considerado (ver Tabla 1).

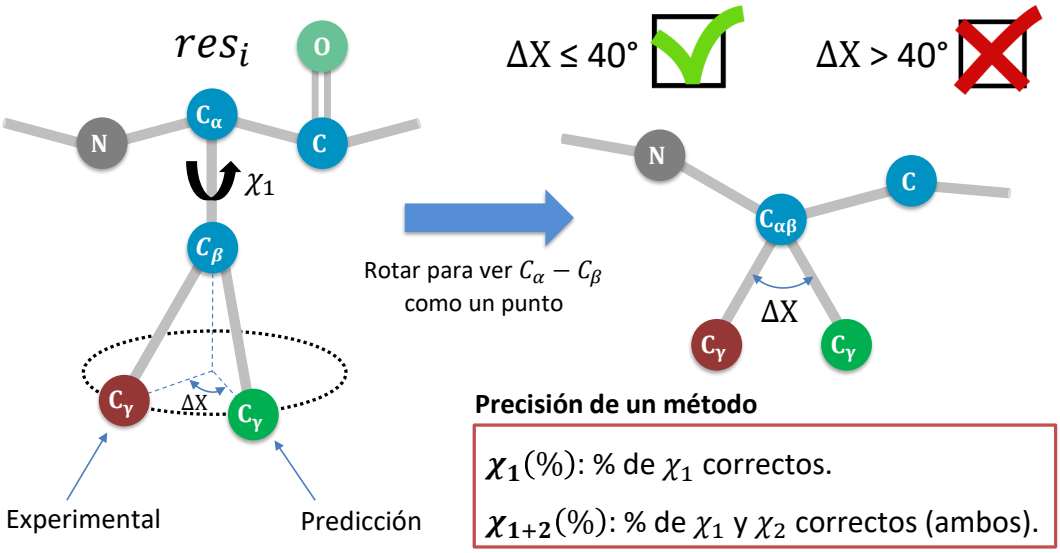

Figura 23. La precisión en una predicción está dada por el porcentaje de ángulos de torsión predichos correctamente. Una predicción es correcta cuando la diferencia de su ángulo de torsión con la de la conformación nativa no supera los 40°.

poca importancia en términos de energía de la proteína (Chandrasekaran y Ramachandran, 1970; Sasisekharan y Ponnuswamy, 1970), lo cual podría ser la razón por la que sólo se consideran $\chi_1(\%)$ y $\chi_{1+2}(\%)$.

En el presente trabajo, para las medidas de $\chi_1(\%)$ y $\chi_{1+2}(\%)$ solamente se consideran los residuos que tienen determinadas todas las posiciones de los átomos de la cadena lateral en la estructura experimental de referencia.

3.6.2. Desviación cuadrática media (RMSD)

El RMSD (*root-mean-square deviation*) se utiliza como una segunda medida de similitud entre estructuras, y su valor es la raíz cuadrada del promedio cuadrático de los valores de distancia entre los átomos correspondientes de las dos estructuras que se comparan. Así:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((x_i - x'_i)^2 + (y_i - y'_i)^2 + (z_i - z'_i)^2\right)} \qquad (5)$$

Donde (x_i, y_i, z_i) representa la posición del átomo i en la estructura de referencia, y (x'_i, y'_i, z'_i) representa la posición del mismo átomo en la estructura predicha. A diferencia del problema de predicción de estructuras, y más aún de la comparación de

dos proteínas distintas, el PSCPP no requiere un alineamiento previo de las dos estructuras; pues tienen la misma cadena principal y se conoce de antemano los pares de átomos equivalentes.

3.6.3. Consideraciones especiales en ciertos tipos de residuos

Casi el 90 % de las estructuras almacenadas en el PDB se obtuvieron mediante cristalografía de rayos X. Recordando el proceso de obtención de una estructura mediante cristalografía de rayos X, explicado en la Subsección 2.2.3.1, las coordenadas tridimensionales de los átomos se determinan a partir de los mapas de densidad electrónica.

Existen ciertos tipos de residuos para los cuales se hacen consideraciones especiales en las métricas empleadas para el PSCCP. En las Figuras 6 y 24 se pueden observar los siguientes casos:

- Ciertos tipos de aminoácidos tienen una estructura simétrica en la cadena lateral (ARG, ASP, GLU, PHE y TYR) (Dunbrack, 2002; Eyal *et al.*, 2004; Cao *et al.*, 2011). Por ejemplo, en la fenilalanina (PHE), la composición del anillo aromático es simétrica considerando los átomos pesados; por lo que existen dos interpretaciones posibles del mapa de densidad electrónica: la conformación finalmente reportada y una conformación que resulta de “dar vuelta” la primera. Esto puede verse en la parte superior de la Figura 24 para el caso del aminoácido PHE.
- En otros tipos de residuos (ASN, HIS y GLN) no se tiene una simetría en la conformación de la cadena lateral, pero de igual manera puede existir una confusión en la interpretación del mapa de densidad electrónica, dándose el mismo caso de tener las dos posibilidades mencionadas en el punto anterior. En la parte inferior de la Figura 24 se muestra el caso de la glutamina (GLN).

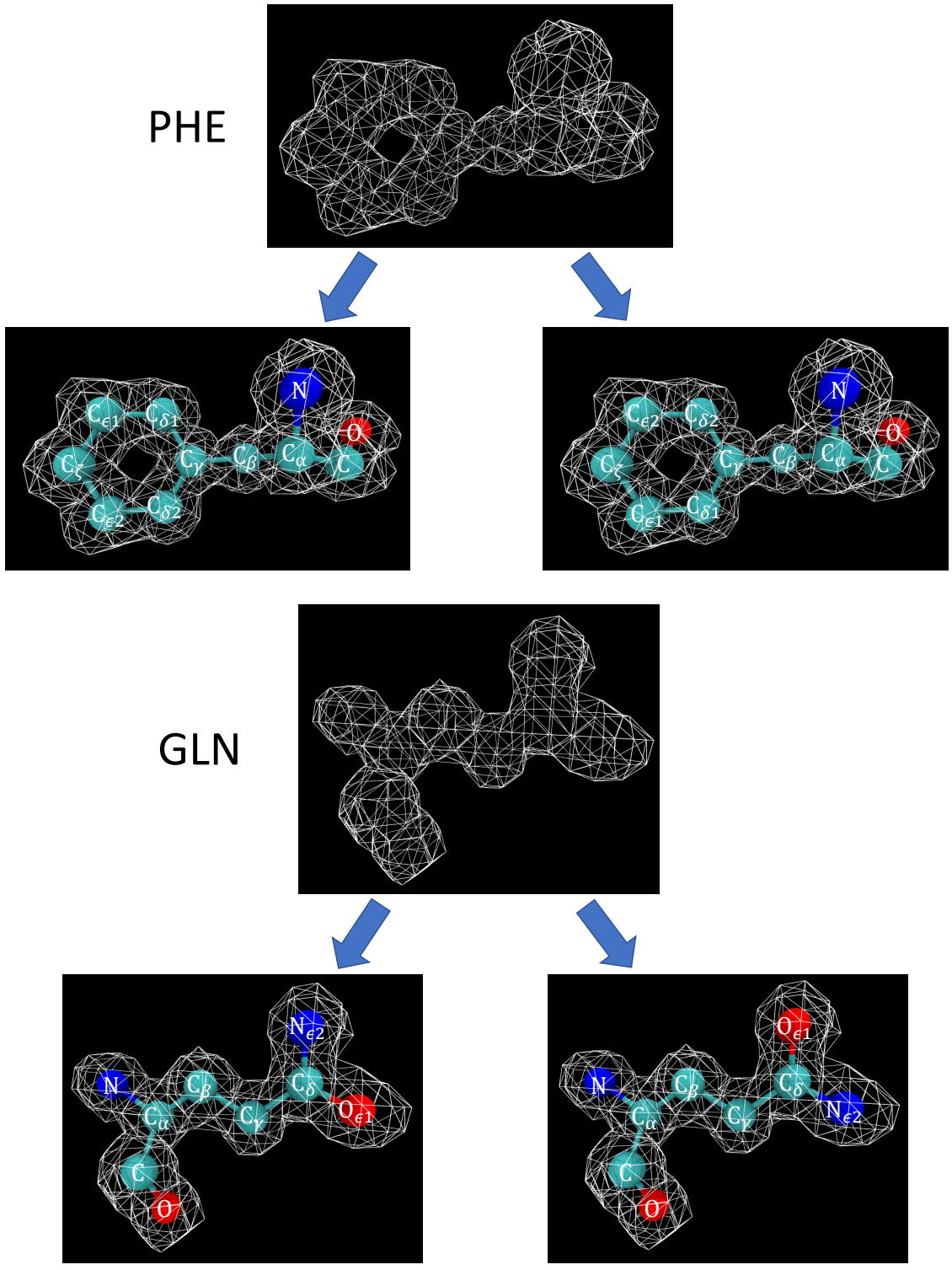

Figura 24. Dos posibles interpretaciones para ciertos tipos de residuos a partir del mapa de densidad electrónica. En el caso de la fenilalanina (PHE), existe una simetría en el anillo. En el caso de la glutamina (GLN) no se da esto (considerando $O_{\epsilon1}$ y $N_{\epsilon2}$), aunque igual existen dos interpretaciones posibles al mapa de densidad electrónica. Imagen obtenida mediante VMD (Humphrey *et al.*, 1996).

Por esta razón, al momento de obtener las métricas de una cierta conformación

Tabla 2. Ángulos de torsión de ciertos tipos de residuo a los que debe sumarse 180° al considerar la precisión y el RMSD.

Ángulo de torsión	Tipos de residuo
χ_2	ASN, ASP, HIS, PHE, TYR
χ_3	GLN, GLU
χ_5^*	ARG

de la cadena lateral se consideran dos posibilidades: la conformación predicha por el método y la conformación “rotada”. El giro se hace sumándole 180° al ángulo de torsión relacionado a la región simétrica. Volviendo al caso de PHE, el giro se hace sumándole 180° a χ_2. En la Tabla 2 se indica cada tipo de residuo particular y el ángulo de torsión al que debe sumarse 180°.

Al considerar las dos conformaciones posibles para estos residuos particulares, el ángulo de torsión más cercano al de la conformación experimental se toma para considerar la precisión; y también se elige el menor valor de RMSD (Eyal *et al.*, 2004; Cao *et al.*, 2011). El caso de la arginina (ARG) es peculiar: debido a su polaridad y a la longitud de su cadena lateral, generalmente es un residuo expuesto al solvente. Esto ocasiona que la misma esté en constante movimiento, y en especial los átomos en el extremo de la cadena lateral. Por esta razón se le asigna un valor fijo de 180° a χ_5 (Corona, 2010), lo cual sólo suele ser relevante para el cálculo del RMSD.

3.6.4. Colisión

Cuando se definió las interacciones de Van der Waals en la Subsección 3.3.2.1, se indicó que a distancias menores a un cierto umbral aparece una fuerza de repulsión que crece rápidamente a medida que decrece la distancia entre los átomos. Esta fuerza de repulsión resulta de la superposición entre las nubes de electrones de ambos átomos. Así aparece el concepto de **radio de Van der Waals**, que se define como el radio de una esfera imaginaria que representa el espacio ocupado por un determinado átomo. Por lo tanto, a distancias menores al radio de Van der Waals empiezan a manifestarse las fuerzas de repulsión.

Otra métrica de calidad importante para el PSCPP es el número de *colisiones* en la estructura predicha. Ocurre una **colisión** entre un par de átomos cuando la distancia entre los mismos es menor que la suma de los respectivos radios de Van der Waals

multiplicada por un factor β. Los valores típicos de β son 0.6 (Lu *et al.*, 2008a) y 0.7 (Cao *et al.*, 2011; Miao *et al.*, 2011). En este trabajo se empleó un valor de β igual a 0.6. Nagata *et al.* (2012) definieron dos clases de colisiones: moderadas (con $\beta = 1.0$) y severas (con $\beta = 0.8325$). Cabe resaltar que valores menores de β cuentan un número más grande de colisiones.

Se consideran ciertas excepciones a la hora de considerar las colisiones entre los átomos de un par de residuos en una estructura:

- Átomos de los enlaces disulfuro entre cisteínas.
- El átomo C_δ de una prolina (PRO) y el átomo C del residuo anterior.
- Dos átomos de oxígeno, en un cierto intervalo de distancias, ya que podrían estar formando un enlace de hidrógeno.

3.7. Métodos del estado del arte

El primer paso del presente trabajo fue evaluar el estado actual de los métodos para el PSCPP, comparando cinco métodos del estado del arte seleccionados: SCWRL4 (Krivov *et al.*, 2009), OPUS-Rota (Lu *et al.*, 2008a), CIS-RR (Cao *et al.*, 2011), RASP (Miao *et al.*, 2011) y SIDEpro (Nagata *et al.*, 2012). La justificación de esta elección es la siguiente: (i) SCWRL4 es el método más completo y más conocido para el PSCPP, lo cual puede verse por su alto número de citas y por ser el método de referencia en la publicación de otros para el PSCPP. (ii) OPUS-Rota sistemáticamente obtiene los mejores valores para $\chi_1(\%)$ y $\chi_{1+2}(\%)$ en varios trabajos publicados (Brizuela *et al.*, 2015; Liang *et al.*, 2011a; Miao *et al.*, 2011). (iii) RASP es el método más rápido, y (iv) CISRR logra el menor número de colisiones. Finalmente, (v) SIDEpro es el método más consistente en todos los criterios, pues más adelante se muestra que siempre está posicionado en el segundo lugar para cada métrica empleada. Todos estos métodos están disponibles públicamente y son fáciles de instalar y ejecutar.

Existen otras propuestas recientes, tales como OSCAR-star (Liang *et al.*, 2011a), SCMF-PDRL (Francis-Lyon y Koehl, 2014), pacoPacker (Quan *et al.*, 2014), BetaSCPWeb (Ryu *et al.*, 2016) y un método de Proteus orientado al diseño de proteínas pero evaluado mediante el PSCPP (Gaillard *et al.*, 2016). Sin embargo, estos métodos no mejoran

el desempeño de los cinco considerados. Además, estos métodos no están disponibles públicamente (a excepción de BetaSCPWeb, pero tiene un desempeño más bajo que los métodos seleccionados).

A continuación se describe cada uno de los métodos elegidos para la comparación de desempeño. Sus funciones de *score* contienen únicamente términos que consideran interacciones internas de residuos y entre pares de residuos.

3.7.1. OPUS-Rota

OPUS-Rota (Lu *et al.*, 2008a) emplea una biblioteca de rotámeros dependiente de la cadena principal (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997). El término de interacciones internas en un residuo está relacionado a las probabilidades de los rotámeros seleccionados en una cierta estructura candidata, el cual se llama *término de frecuencia de rotámeros*. Las interacciones de Van der Waals se modelan en una componente de las interacciones entre pares de residuos. Estos dos términos se emplean en casi todos los métodos para el PSCPP. OPUS-Rota también incorpora dos términos únicos en las interacciones entre pares: un potencial dependiente de la orientación de las cadenas laterales y un término que aproxima las interacciones con el solvente.

El método de búsqueda de OPUS-Rota se basa en el recocido simulado[8] con el empleo del baño de calor de Monte Carlo (Aarts y Lenstra, 2003). Para cada iteración, todos los rotámeros de cada residuo se consideran una sola vez a una temperatura constante. Luego, la temperatura disminuye gradualmente desde 2.5 a 0.05 a lo largo de 97 iteraciones, seguida de tres iteraciones a una temperatura igual a 0 (lo que corresponde a tres iteraciones de una búsqueda local).

3.7.2. SCWRL4

SCWRL4 (Krivov *et al.*, 2009) es uno de los mejores métodos para el empacamiento de la cadena lateral (Gaillard *et al.*, 2016), y es uno de los métodos más empleados para dar soluciones aproximadas al PSCPP. Emplea una de las más recientes bibliotecas de rotámeros dependientes de la cadena principal (Shapovalov y Dunbrack, 2011).

[8] Más adelante se tratará con más detalle sobre el recocido simulado y otros algoritmos basados en la búsqueda local.

Este método considera tanto un modelo de rotámeros rígidos (RRM, por sus siglas en inglés) como un modelo de rotámeros flexibles (FRM, por sus siglas en inglés), donde se permite ciertas variaciones en los ángulos de torsión de los rotámeros de la biblioteca en función a las desviaciones estándar que aparecen en la misma. En el RRM, el término de interacciones internas considera la probabilidad del rotámero en función de la máxima probabilidad de los rotámeros posibles para el residuo. Los términos de interacciones entre pares de residuos consideran las interacciones de Van der Waals y enlaces de hidrógeno.

Se usa un método determinístico para la búsqueda de soluciones, donde las interacciones entre residuos se representan mediante un grafo. Así, la búsqueda se realiza mediante la descomposición de aristas, la aplicación del algoritmo DEE[9] (Desmet *et al.*, 1992) y la descomposición de árbol. Un punto importante es que SCWRL4 es el único método entre los seleccionados con una opción para la consideración de los contactos simétricos[10] en la predicción de la conformación de la cadena lateral de la proteína.

3.7.3. CIS-RR

CIS-RR (Cao *et al.*, 2011) emplea la misma biblioteca de rotámeros que OPUS-Rota (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997). La función de score se adaptó de SCWRL3 (Canutescu *et al.*, 2003) y consiste de: (i) un término en función de la probabilidad del rotámero, (ii) un término empírico de Van der Waals, y (iii) un término para los enlaces disulfuro. Estos últimos dos términos se basan en los de SCWRL3.

El método de búsqueda de CIS-RR se enfoca en minimizar las colisiones atómicas en la estructura predicha. Ésto se realiza introduciendo una fase llamada *relajación de rotámeros* (RR), la cual ha demostrado disminuir significativamente el número de colisiones.

La conformación inicial de cada residuo se construye con los rotámeros con la mayor probabilidad en cada posición. Luego, para cada residuo *i*, cada rotámero se optimiza mediante RR y se evalúa en términos de colisiones con otros residuos que se

[9] Dead-end elimination: se eliminan rotámeros que no pertenecen a la estructura de mínimo score.
[10] Definido en la Subsección 2.2.3.1

mantienen fijos. Si un rotámero *r* colisiona con otros residuos, el mismo se mantiene fijo temporalmente, y los residuos con los que colisiona cambian a cada uno de sus rotámeros optimizados por la fase RR. Después de que todos los rotámeros del residuo *i* se han explorado, las conformaciones de las cadenas laterales de *i* y los residuos con los que colisiona se actualizan con los ángulos de torsión que generen la menor cantidad de colisiones durante la búsqueda. Este proceso se repite desde los residuos menos expuestos a los más expuestos hasta que converge a un valor estable (Cao *et al.*, 2011).

3.7.4. RASP

RASP (Miao *et al.*, 2011) también emplea la misma biblioteca que OPUS-Rota (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997). El término de interacciones internas es el mismo que el de SCWRL4, y las interacciones entre pares de residuos considerados son las de Van der Waals, enlaces disulfuro y enlaces de hidrógeno. El término para las interacciones de Van der Waals es una adaptación al empleado en OPUS-Rota y el término para los enlaces disulfuro se basa en el de SCWRL3 (Canutescu *et al.*, 2003).

En relación al método de búsqueda, RASP sigue el mismo enfoque que SCWRL4 con algunas modificaciones. El algoritmo DEE (Desmet *et al.*, 1992) se utiliza al inicio para reducir el tamaño del espacio de búsqueda. Luego, se construye un grafo de interacciones y se divide en componentes biconectados (Tarjan, 1972). Para grafos de interacciones de gran tamaño, se realiza un recocido simulado de Monte Carlo (Aarts y Lenstra, 2003). Para casos de grafos más pequeños se aplica una variante de la estrategia de ramificación y acotamiento (*branch-and-bound*), denominada ramificación y terminación (*branch-and-terminate*) (Gordon y Mayo, 1999).

RASP emplea un algoritmo de optimización guiado por la detección de colisiones, con el fin de disminuir las colisiones atómicas debido al uso de rotámeros fijos. Este método se enfoca en lo siguiente: (i) la rápida generación de estructuras iniciales de alta calidad, y (ii) la rápida eliminación de colisiones atómicas mediante la relajación de los residuos involucrados en ellas (Miao *et al.*, 2011).

3.7.5. SIDEpro

SIDEpro (Nagata *et al.*, 2012) también usa la misma biblioteca de rotámeros que OPUS-Rota (Dunbrack y Karplus, 1993; Dunbrack y Cohen, 1997). Este método emplea una familia de redes neuronales para calcular la función de score. Cada red neuronal se especializa en calcular un término particular del score total asociado con un tipo de aminoácido y con la distancia entre pares de átomos específicos. A pesar de no disponer de objetivos con energía definida que puedan servir para la etapa de entrenamiento de la función de score, las redes neuronales se pueden optimizar empleando la información de las estructuras en el PDB y convirtiendo las energías de las redes neuronales a probabilidades; optimizando estas probabilidades usando métodos de Monte Carlo basados en cadenas de Markov (Geyer, 1992).

Luego de excluir todos los rotámeros que causan colisiones atómicas, SIDEpro realiza predicciones asignando inicialmente probabilidades a cada uno de los rotámeros de la biblioteca para cada residuo. Luego, el predictor actualiza estas probabilidades mediante las redes neuronales entrenadas. Al lograr una convergencia en los valores de probabilidad, las conformaciones de las cadenas laterales se determinan por rotámeros de mayor probabilidad para cada residuo. Finalmente, se aplica una fase de reducción de colisiones para obtener la estructura predicha (Nagata *et al.*, 2012).

3.7.6. Comparaciones anteriores entre métodos del estado del arte

Cuando se propone un nuevo método para el PSCPP, generalmente se compara con algún otro método del estado del arte. Existen también trabajos que se enfocan en la comparación de los mismos bajo diferentes criterios. La precisión en la predicción es la principal métrica empleada para evaluar el desempeño de un método en particular. La Tabla 3 presenta un resumen de los valores reportados en la literatura.

Como puede verse, se utilizan varios conjuntos de proteínas para las evaluaciones; por lo que las precisiones obtenidas son diferentes entre distintos trabajos. Algo que llama la atención es que existen diferentes precisiones reportadas para un mismo método en un mismo conjunto de pruebas (por ejemplo, ver el caso de SCWRL4), lo cual puede deberse a las siguientes razones:

Tabla 3. Precisiones reportadas sobre OPUS-Rota, SCWRL4, CIS-RR, RASP y SIDEpro para distintos conjuntos de prueba.

Método	Autores	Nro. de proteínas	χ_1(%)	χ_{1+2}(%)
OPUS-Rota	Lu *et al.* (2008a)	65	89.00	79.10
	Liang *et al.* (2011a)	218	86.60	75.70
	Miao *et al.* (2011)	379	85.03	75.05
	Francis-Lyon y Koehl (2014)	831	81.69	66.17
SCWRL4	Krivov *et al.* (2009)	379	86.00	75.00
	Cao *et al.* (2011)	65	85.80	76.30
	Liang *et al.* (2011a)	218	85.10	74.00
	Miao *et al.* (2011)	379	85.03	75.44
	Nagata *et al.* (2012)	379	85.43	73.47
	Peterson *et al.* (2014)	231	85.20	72.00
	Francis-Lyon y Koehl (2014)	831	79.61	63.94
	Gaillard *et al.* (2016)	18	89.00	78.00
CIS-RR	Cao *et al.* (2011)	65	86.40	76.70
	Liang *et al.* (2011a)	218	84.70	73.10
	Miao *et al.* (2011)	379	84.88	74.88
RASP	Miao *et al.* (2011)	379	85.10	74.71
	Peterson *et al.* (2014)	231	85.20	71.00
SIDEpro	Nagata *et al.* (2012)	379	86.14	74.15

- Algunos trabajos reportan el promedio de las precisiones para cada proteína, mientras que otros calculan la precisión sobre el total de residuos en el conjunto de proteínas.
- Algunos trabajos descartan los residuos con átomos faltantes, mientras que otros consideran todos los ángulos de torsión que puedan obtenerse con los átomos disponibles en la estructura de referencia.
- La consideración de los tipos de residuos especiales señalados en la Subsección 3.6.3.
- En el caso de las proteínas multiméricas[11] con subunidades iguales (llamadas homo-multiméricas), se considera una sola cadena polipeptídica.
- La consideración o no de los residuos con más de una conformación y la elección de la conformación de referencia.
- Consideración de los residuos de acuerdo a un valor umbral en el percentil de los valores de densidad electrónica (Krivov *et al.*, 2009).

[11] Proteínas con más de una cadena polipeptídica.

Por lo tanto, para realizar la comparación de los métodos para el PSCPP en el presente trabajo, primero se definen los conjuntos de prueba que se utilizarán.

3.7.7. Conjuntos de prueba o “datasets”

Para cada experimento realizado en este trabajo de investigación se emplearon cuatro conjuntos de proteínas. Las mismas están compuestas por 65, 360, 693 y 2230 proteínas, respectivamente; y todas las estructuras de estas proteínas se obtuvieron experimentalmente mediante cristalografía de rayos X. En el Anexo A.2 se provee información acerca de la lista de los identificadores PDB de las proteínas en cada conjunto de prueba. Tres de estos conjuntos de prueba se basan en conjuntos utilizados previamente, mientras que el restante es una adaptación del propuesto en una tesis de maestría de CICESE (Corona, 2010).

Más adelante se verá el efecto de considerar los *contactos simétricos* en las estructuras cristalográficas, los cuales se definen en la Subsección 2.2.3.1. Un punto importante es que ninguno de los trabajos anteriores (a excepción de la propuesta de SCWRL4) considera los contactos simétricos para evaluar el desempeño de los métodos para el PSCPP, a pesar de emplear estructuras obtenidas mediante cristalografía de rayos X. Estos contactos son muy relevantes para la consideración de las interacciones entre las cadenas laterales de residuos expuestos al solvente y las demás copias de la proteína en el cristal (Dunbrack y Karplus, 1993). De manera a obtener los contactos simétricos a partir de las estructuras almacenadas en el PDB, se empleó la información del sitio web de WHATIF[12] (Hekkelman *et al.*, 2010). A continuación se da una breve descripción de cada conjunto de prueba:

- Dataset-65: éste es el conjunto empleado para evaluar el método NCN (Peterson *et al.*, 2004) para el PSCPP, y es el mismo usado para evaluar OPUS-Rota (Lu *et al.*, 2008a). Dentro de este conjunto, un subconjunto de 30 proteínas se seleccionó a partir de otro conjunto de prueba (Liang y Grishin, 2002). Para este subconjunto, el límite de similitud de secuencias fue de 50 %, el límite de resolución fue de 1.8 Å y el factor R fue de 20 %. Se seleccionaron únicamente las proteínas monoméricas (de una sola cadena) con longitudes entre 100 y 500 re-

[12] http://swift.cmbi.ru.nl/gv/lists/

siduos, sin átomos faltantes en la cadena lateral y sin ligandos. Otro subconjunto de 28 proteínas se seleccionó del trabajo previo de Xiang y Honig (2001), donde algunas tienen una resolución entre 0.83 Å y 1.4 Å con una identidad de pares de secuencias menor que 20 % y las demás tienen más de 40 residuos y una resolución mejor que 1.2 Å. Las siete proteínas restantes se seleccionaron del PDB de acuerdo a los siguientes criterios: una resolución mejor que 1.2 Å y una longitud de secuencia entre 150 y 300 residuos. No existen colisiones entre los átomos de las estructuras en el Dataset-65, y todas ellas están disponibles en el sitio web de WHATIF.

- Dataset-360: Este es un subconjunto del conjunto de 379 proteínas propuesto para evaluar SCWRL4 (Krivov *et al.*, 2009). Las proteínas tienen una longitud de secuencia entre 40 y 1000 residuos, con una resolución mejor que 1.8 Å, una similitud máxima entre pares de secuencia de 30 % y un factor R máximo de 20 %. Existen ocho colisiones atómicas dentro del Dataset-360. Un total de 19 proteínas no se encuentran en el sitio web de WHATIF, por lo que se tienen 360 estructuras en este conjunto.

- Dataset-693: Este es un subconjunto del conjunto de 721 proteínas propuesto por (Corona, 2010), que fue posteriormente publicado (Colbes *et al.*, 2016). Las proteínas que forman parte de este conjunto son monoméricas, con una longitud de secuencia entre 40 y 400 residuos, un único dominio bajo la clasificación SCOP[13] (dentro de las clases a,b,c y d) (Murzin *et al.*, 1995), un factor R máximo de 20 %, una resolución mejor que 2 Å y una identidad máxima de secuencia de 25 %. Este conjunto se propuso principalmente para evitar sesgos hacia cualquiera de los métodos evaluados. Existen 28 colisiones atómicas dentro del Dataset-693. Un total de 28 proteínas no se encuentran en el sitio web de WHATIF, por lo que se tienen 693 estructuras experimentales en este conjunto.

- Dataset-2230: Este es un subconjunto del conjunto propuesto para evaluar RASP (Miao *et al.*, 2011), el cual consiste de 2412 proteínas. Estas proteínas, obtenidas del servidor PISCES (Wang y Dunbrack, 2003), tienen una resolución mejor que 1.8 Å, un factor R máximo de 25 % y una identidad máxima de secuencia de 25 %. Existen 132 colisiones atómicas dentro del Dataset-2230. Un total de 28

[13] Clasificación de las proteínas de acuerdo a las características de sus estructuras.

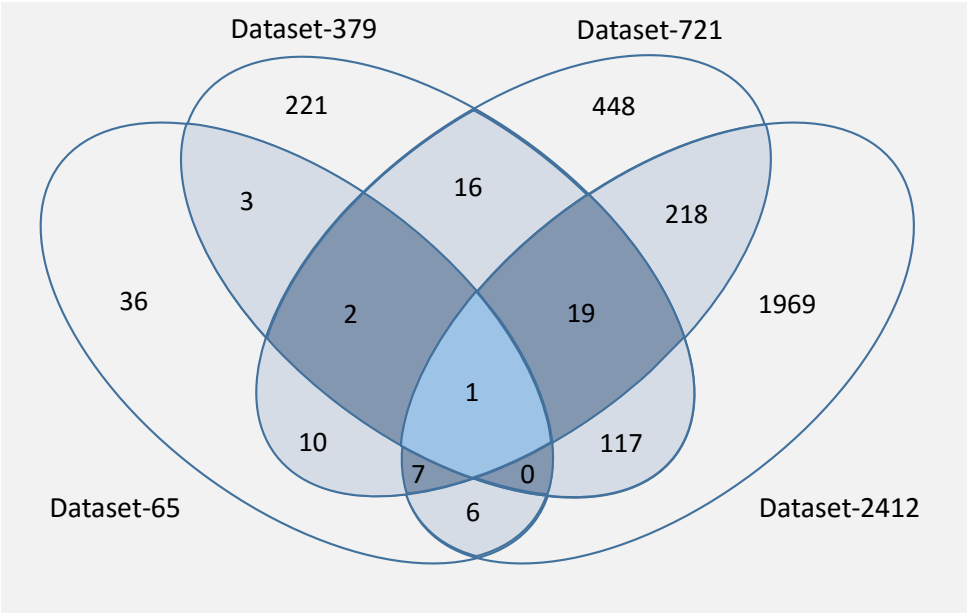

Figura 25. Relación entre los cuatro conjuntos de proteínas empleados para los experimentos.

proteínas no se encuentran en el sitio web de WHATIF, por lo que se tienen 2230 estructuras experimentales en este conjunto.

El diagrama de Venn de la Figura 25 muestra las coincidencias entre los conjunto de prueba definidos. Todas estas estructuras se encuentran disponibles públicamente[14].

3.7.8. Resultados derivados de las comparaciones

Los criterios seguidos para calcular la precisión y el RMSD de una estructura, en todos los experimentos del presente trabajo de investigación, fueron los siguientes:

- Los valores de $\chi_1(\%)$ y $\chi_{1+2}(\%)$ se obtienen considerando la totalidad de los residuos del conjunto de prueba utilizado.
- Un residuo se considera solamente si en el archivo PDB de la estructura experimental de referencia aparecen todos los átomos de su cadena lateral.
- En el caso de residuos con múltiples conformaciones en el archivo PDB, se elige la conformación de mayor ocupación.

[14] https://figshare.com/articles/Datasets_for_testing_the_performance_of_a_method_for_PSCPP/3483089/1

- En el caso de proteínas con múltiples cadenas, sólo se considera la primera.
- Se consideran los tipos de residuos especiales, definidos en la Subsección 3.6.3.

Se consideran tres tipos de residuos:

1. *Enterrado*: si el porcentaje de acceso al solvente del residuo es menor a 20 %.
2. *Expuesto*: si el porcentaje de acceso al solvente del residuo es mayor o igual a 20 %, pero no es un contacto simétrico.
3. *Simétrico*: si el residuo es un contacto simétrico y no es del tipo enterrado.

El porcentaje de acceso al solvente de cada residuo se calculó mediante un script en Python que utiliza al programa PyMOL (Schrödinger, LLC, 2015) y el criterio de Fraczkiewicz y Braun (1998). También puede calcularse mediante el programa DSSP (Touw *et al.*, 2014). Todos los experimentos se realizaron en un equipo con Intel Core i7-4500U a 1.8GHz, con 12GB de RAM y con los sistemas operativos recomendados por los autores de los métodos considerados: Windows 10 para SCWRL4 y Ubuntu 14.04 para los demás.

De las comparaciones anteriores, mostradas en la Tabla 3, puede verse que los métodos considerados tienen una precisión de aproximadamente 85 % para χ_1(%) y 75 % para χ_{1+2}(%). Los resultados de la comparación realizada en este trabajo se muestran en las tablas 4 y 5, teniéndose un promedio de aproximadamente 87 % para χ_1(%) y 77 % para χ_{1+2}(%). Puede observarse que la consideración de los contactos simétricos eleva la precisión de los métodos considerados. Esto es de esperarse, ya que las estructuras experimentales de referencia están influenciadas por las interacciones con proteínas vecinas; y esa misma situación se da en la predicción al considerar los residuos de otras proteínas vecinas que interactúan con la estructura objetivo.

Tabla 4. Resultados para cada método en cada conjunto de prueba.

Precisión total y número de colisiones

	Dataset-65 (Col.: 0)			Dataset-360 (Col.: 8)			Dataset-693 (Col.: 28)			Dataset-2230 (Col.: 132)		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
SCWRL4	86.45	77.18	76	86.45	77.05	498	87.05	77.9	929	86.69	77.57	3855
OPUS-Rota	**88.75**	**80.46**	114	**87.49**	**78.2**	380	**88.38**	**79.38**	1108	**88.03**	**79.01**	4133
CIS-RR	87.07	78.07	**13**	86.11	76.47	**48**	86.66	77.16	**143**	86.39	76.87	**554**
RASP	86.63	75.54	154	85.86	75.09	1007	86.36	75.9	2442	86.58	76.17	9105
SIDEpro	87.97	79.02	16	86.77	77.65	129	87.53	78.6	323	87.24	78.43	1191

Precisión - Residuos enterrados

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	94.28	90.98	94.95	90.56	95.55	**91.67**	94.93	91.72
OPUS-Rota	**95.94**	92.29	**95.71**	90.92	**96.43**	91.49	**96.13**	91.68
CIS-RR	95.37	**92.53**	95.46	**91.67**	95.55	**91.67**	94.97	**91.82**
RASP	93.92	88.49	94.47	88.32	94.44	88.78	94.36	88.97
SIDEpro	95.37	91.1	95.33	90.14	95.46	90.76	95.08	90.76

Precisión - Residuos expuestos

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	86.63	77.59	86.14	77.03	87.23	78.49	86.62	77.69
OPUS-Rota	**88.93**	**81.19**	**87.31**	**78.49**	**88.55**	**80.19**	**88.06**	**79.37**
CIS-RR	87.25	78.72	86	76.62	86.89	78.05	86.37	77.09
RASP	87.03	76.57	85.91	75.6	86.77	76.85	86.68	76.6
SIDEpro	87.82	79.49	86.7	78.15	87.78	79.49	87.27	78.75

Precisión - Residuos simétricos

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	83.11	72.41	83.92	73.54	84.28	74.04	83.65	73.25
OPUS-Rota	**85.62**	**75.83**	**84.85**	**74.46**	**85.77**	**75.64**	**84.86**	**74.66**
CIS-RR	83.54	72.77	83	72.29	83.72	72.77	83.12	72.14
RASP	83.2	70.19	82.75	70.93	83.44	71.8	83.37	71.64
SIDEpro	85.25	74.7	83.87	73.69	84.85	74.75	84.15	74.27

Tabla 5. Resultados para cada método en cada conjunto de prueba (continuación).

RMSD (promedio y desviación estándar) y tiempo total de ejecución

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	RMSD	Tiempo	RMSD	Tiempo	RMSD	Tiempo	RMSD	Tiempo
SCWRL4	1.536 ± 0.21	5m 26s	1.586 ± 0.24	36m 55s	1.523 ± 0.24	64m 22s	1.582 ± 0.29	213m 57s
OPUS-Rota	**1.354 ± 0.23**	5m 8s	**1.496 ± 0.24**	28m 32s	**1.432 ± 0.23**	53m 56s	**1.487 ± 0.27**	193m 58s
CIS-RR	1.44 ± 0.26	8m	1.581 ± 0.26	51m 19s	1.521 ± 0.26	95m 38s	1.577 ± 0.29	380m 48s
RASP	1.608 ± 0.24	**10s**	1.64 ± 0.26	**1m 9s**	1.592 ± 0.26	**2m 8s**	1.63 ± 0.29	**9m 4s**
SIDEpro	1.441 ± 0.23	1m 14s	1.542 ± 0.24	7m 6s	1.476 ± 0.24	13m 56s	1.533 ± 0.3	49m 55s

En general, los cinco métodos considerados tienen desempeños similares, en el sentido de que la mayoría destaca en alguna métrica: (i) OPUS-Rota tiene sistemáticamente los mejores valores de χ_1(%), χ_{1+2}(%) y RMSD; (ii) CIS-RR tiene el menor número de colisiones; (iii) RASP es el más rápido; y (iv) SIDEpro se encuentra en segundo lugar en cada métrica de calidad.

En relación a los tipos de residuos, puede notarse que los residuos enterrados permiten una precisión más alta que los expuestos: aproximadamente 8 % para χ_1(%) y 13 % para χ_{1+2}(%). Esto se debe principalmente al mayor número de restricciones *estéricas*[15] que tienen estos residuos en relación a los expuestos. Además, los residuos con mayor flexibilidad pueden estar incorrectamente determinados en la estructura cristalográfica y estos residuos generalmente son expuestos. Estos residuos son sensibles al solvente que los rodea, por lo que es de esperarse que los métodos que incluyan términos que consideren las interacciones con el solvente tengan un mejor desempeño. En la Tabla 4 puede verse que éste es el caso de OPUS-Rota, el cual considera estas interacciones y tiene la mejor precisión.

Los factores experimentales, como la resolución y el factor de temperatura (definidos en la Subsección 2.2.3.1), pueden afectar la precisión; y por esta razón se hicieron experimentos para analizar la influencia de estos factores en los valores de precisión. Así como lo hicieron Carugo y Argos (1997), se normalizó los factores de temperatura antes de comparar los valores entre diferentes estructuras cristalográficas. Los resultados en la Tabla 6 muestran que, si no se consideran los residuos con factores de temperatura en el peor cuartil, la precisión aumenta en aproximadamente 4 % para χ_1(%) y 6 % para χ_{1+2}(%). La influencia de la resolución de las estructuras cristalográficas de referencia es menor: la diferencia entre los dos grupos (divididos de acuerdo al valor de la mediana para cada conjunto) es de aproximadamente 1.5 % para χ_1(%) y 3 % para χ_{1+2}(%), como puede verse en la Tabla 7.

No se realizó un análisis estadístico más detallado debido a que el objetivo principal de estas pruebas es mostrar el aparente estancamiento de los resultados de precisión para el PSCPP. La interrogante que surge es si este promedio de precisión de los métodos - alrededor de 87 % para χ_1(%) y 77 % para χ_{1+2}(%) - son los valores máximos

[15] Relacionadas al espacio que ocupa un átomo, dado por la esfera determinada por el radio de Van der Waals.

alcanzables para el PSCPP. Para contestar esta pregunta, se debe examinar cada uno de los tres componentes principales de los métodos para el PSCPP.

3.8. Máxima precisión alcanzable

La mayoría de los métodos del estado del arte para el PSCPP trabaja con una biblioteca de rotámeros dependiente de la cadena principal (Krivov *et al.*, 2009; Lu *et al.*, 2008a; Cao *et al.*, 2011; Miao *et al.*, 2011; Nagata *et al.*, 2012; Quan *et al.*, 2014; Francis-Lyon y Koehl, 2014). La razón principal de ésto es la contribución de los ángulos de torsión y frecuencias de los rotámeros en función a la conformación de la cadena principal, al proporcionar y favorecer las conformaciones de la cadena lateral que estén más relacionadas con la cadena principal de un caso del PSCPP. Para evaluar la influencia de una biblioteca de rotámeros, se calculó la máxima precisión alcanzable a través de la biblioteca de rotámeros dependiente de la cadena principal de Dunbrack y Cohen (1997) (versión de Mayo de 2002). Esta biblioteca se usa en cuatro de los cinco métodos del estado del arte considerados en este trabajo.

Como se dijo anteriormente, la biblioteca de rotámeros define el espacio de búsqueda. Si las conformaciones de los rotámeros en ella están muy alejadas de las conformaciones experimentales de las estructuras de proteínas de entrada, entonces no se podrán obtener resultados precisos por más que la función de score y el algoritmo de búsqueda sean los ideales. Por lo tanto, resulta conveniente calcular la máxima precisión a la que se puede aspirar con una biblioteca de rotámeros determinada.

Tabla 6. Resultados para cada método y cada conjunto de prueba, considerando los residuos con cadenas laterales con factor de temperatura en el percentil 75.

Precisión total

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	90.15	83.8	90.3	83.49	90.71	84.31	90.03	83.28
OPUS-Rota	92.16	86.53	91.2	84.45	91.84	85.15	91.25	84.44
CIS-RR	90.73	84.58	89.76	82.67	90.09	83.12	89.56	82.37
RASP	89.94	81.74	89.39	81.14	89.7	81.63	89.64	81.4
SIDEpro	91.53	85.67	90.52	83.84	90.94	84.47	90.37	83.73

Tabla 7. Resultados para cada método y cada conjunto de prueba, considerando los residuos de las estructuras de acuerdo a la resolución de la misma. El límite de separación entre los dos grupos está dado por la mediana de los valores de resolución en cada conjunto de prueba.

Precisión total - Mejores resoluciones

	Dataset-65 Res.: 0.78 Å- 1.20 Å		Dataset-360 Res.: 0.96 Å- 1.60 Å		Dataset-693 Res.: 0.62 Å- 1.60 Å		Dataset-2230 Res.: 0.62 Å- 1.55 Å	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	87.83	78.89	86.99	78.28	87.97	79.37	87.37	78.68
OPUSRota	90.04	82.63	88.04	79.23	89.32	80.92	88.76	80.18
CIS-RR	88.43	79.64	86.35	77.10	87.50	78.47	87.06	77.91
RASP	88.38	77.83	86.44	76.26	87.28	77.44	87.22	77.26
SIDEpro	89.51	80.98	87.29	78.63	88.51	80.11	87.92	79.54

Precisión total - Peores resoluciones

	Dataset-65 Res.: 1.20 Å- 1.80 Å		Dataset-360 Res.: 1.60 Å- 1.80 Å		Dataset-693 Res.: 1.60 Å- 2.00 Å		Dataset-2230 Res.: 1.55 Å- 1.80 Å	
	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)	χ_1(%)	χ_{1+2}(%)
SCWRL4	85.35	75.91	85.95	75.91	86.20	76.57	86.04	76.52
OPUSRota	87.71	78.83	86.98	77.24	87.51	77.98	87.34	77.91
CIS-RR	86.00	76.90	85.89	75.90	85.89	75.98	85.75	75.89
RASP	85.25	73.84	85.32	74.03	85.53	74.50	85.96	75.15
SIDEpro	86.74	77.57	86.30	76.76	86.63	77.24	86.58	77.39

El proceso de cálculo de la máxima precisión alcanzable es el siguiente: para cada residuo de la proteína se selecciona el rotámero más *cercano* de la biblioteca. Este rotámero se define como el que tiene el mayor número de ángulos de torsión de la cadena lateral correctos[16]. Si dos o más rotámeros cumplen con este criterio, se selecciona el que tiene el menor error total (la suma de los errores en todos los ángulos de torsión del residuo en relación a la conformación de la estructura de referencia). Finalmente se calculan los valores de $\chi_1(\%)$ y $\chi_{1+2}(\%)$. La estructura creada de esta forma se llama *Best*[17]. Un proceso similar podría realizarse obteniendo el menor valor de RMSD en cada residuo en relación a la estructura experimental (*Best-RMSD*); aunque en la Tabla 8 se muestran resultados muy parecidos en cuanto a precisión y RMSD empleando ambos criterios. Como es de esperarse, *Best* obtiene, ligeramente, una mejor precisión; y *Best-RMSD* obtiene un RMSD ligeramente menor. La diferencia principal se da en el menor número de colisiones en las estructuras *Best-RMSD*.

Los ángulos de torsión de la cadena lateral guían la construcción de la estructura *Best*, con el objetivo de maximizar la precisión de la estructura creada. Sin embargo, se puede observar que al realizar esta construcción aparece un número considerable de colisiones atómicas. A pesar de que el número de colisiones en *Best-RMSD* es menor, igual sigue siendo mucho mayor que el de una estructura experimental. Por lo tanto, el siguiente paso fue determinar si es posible reducir el número de colisiones en las estructuras *Best* a los niveles alcanzados por los métodos del estado del arte, sin un deterioro significativo en las precisiones alcanzadas por estas estructuras. Para este propósito se ha modificado el algoritmo de búsqueda de CIS-RR (Cao *et al.*, 2011): básicamente se removió el procedimiento de relajación de rotámeros y se usó el número de colisiones atómicas en la estructura como función de score. Esta versión modificada se describe en el Algoritmo 1.

Se usó *Best* como la estructura de entrada para la versión modificada de CIS-RR (es decir, $S_{in} = Best$), y la estructura obtenida con este proceso se llama *Best+* (es decir, $Best+ = S_{CIS}$). La Tabla 9 muestra la máxima precisión alcanzable para las estructuras *Best* y *Best+*. Considerando que métodos como CIS-RR y RASP no usan la totalidad de los rotámeros de la biblioteca, también se calculó la máxima precisión de

[16] De acuerdo al criterio de los 40° de diferencia en los ángulos de torsión.

[17] Porque básicamente es la estructura más similar a la conformación experimental que puede construirse con los rotámeros de la biblioteca.

Algoritmo 1 Versión modificada de CIS-RR para la reducción de colisiones.

Entrada: Estructura S_{Best}.
Salida: Estructura S_{CIS}.

```
1: S_CIS = S_Best
2: min_col = colisiones(S_CIS)
3: repetir
4:   fin = Verdadero
5:   para cada residuo i hacer
6:     para cada rotámero r de i hacer
7:       si r colisiona con otros residuos entonces
8:         para cada residuo j en colisión hacer
9:           para cada rotámero s de j hacer
10:            Asignar temporalmente r a i y s a j, obteniendo S_rs
11:            num_colisiones = colisiones(S_RS)
12:            si num_colisiones < min_col entonces
13:              min_col = num_colisiones
14:              Almacenar r y s
15:              fin = Falso
16:            fin si
17:          fin para
18:        fin para
19:      fin si
20:    fin para
21:    Actualizar S_CIS con los rotámeros (r, s) en los residuos (i, j)
22:  fin para
23: hasta que fin == Verdadero
24: devolver S_CIS
```

sus bibliotecas reducidas de rotámeros de acuerdo a los siguientes criterios:

- CIS-RR: se remueven los rotámeros que tienen asignados una probabilidad menor al 1 %.
- RASP: se ordenan los rotámeros de manera descendente en el valor de la probabilidad (para cada combinación del tipo de residuo y ángulos de torsión de la cadena principal), y se selecciona el grupo de rotámeros con una probabilidad acumulada del 98 %.

Se puede observar que las estructuras *Best* alcanzan valores por encima de 99 % para χ_1(%) y por encima de 97 % para χ_{1+2}(%). Las estructuras *Best*+ tienen valores de precisión ligeramente menores (>97 % para χ_1(%) y >94 % para χ_{1+2}(%)), pero el número de colisiones decrece significativamente, llegando a los niveles alcanzados por los métodos del estado del arte (ver la Tabla 4). De la Tabla 8 puede notarse que

Best-RMSD obtiene menores valores de RMSD y colisiones que *Best*. Pero por otro lado, cuando se reduce el número de colisiones de *Best* para obtener *Best+*, el RMSD aumenta; lo cual indica que un menor valor de RMSD no garantiza, en general, un menor número de colisiones.

Al comparar los resultados de precisión en las Tablas 4 y 9, se puede observar que existe una diferencia del 10 % para χ_1(%) y del 17 % para χ_{1+2}(%) entre la precisión máxima alcanzable (*Best+*) y los resultados de los métodos del estado del arte.

Como se mencionó anteriormente, los componentes de la mayoría de los métodos para el PSCPP son una biblioteca de rotámeros, una función de score y un algoritmo de búsqueda. Los resultados presentados en la Tabla 9 indican que una biblioteca de rotámeros usualmente empleada es capaz de obtener una precisión casi ideal en los conjuntos de prueba del presente trabajo, y este hecho sugiere que las bibliotecas de rotámeros no son la causa de las limitaciones de los métodos del estado del arte para el PSCPP. Entonces surge la pregunta: ¿Son las funciones de score o los algoritmos de búsqueda los principales responsables de las limitaciones de los métodos actuales para el PSCPP? Trabajos recientes para el diseño de proteínas (Li *et al.*, 2013; Liu y Chen, 2016) indican que las funciones de score todavía fallan en modelar correctamente las interacciones dentro de la proteína y de ésta con el medio circundante, por lo que podrían ser las principales responsables de las mejoras marginales en el PSCPP (lo cual conjeturaron Liang y Grishin (2002)). Das (2011) llegó a la misma conclusión al analizar la función de score de Rosetta (Simons *et al.*, 1999) para el problema de predicción de estructura de proteínas pequeñas. Como estos trabajos sugieren la necesidad de enfocarse en mejorar las funciones de score, en el siguiente capítulo se propone un nuevo método para la evaluación de dichas funciones para el PSCPP.

Tabla 8. Resultados para las estructuras más cercanas a la experimental en términos de precisión (*Best*) y en términos del RMSD de la cadena lateral (*Best-RMSD*), usando la biblioteca de rotámeros dependiente de la cadena principal de Dunbrack y Cohen (1997).

Precisión total y número de colisiones

	Dataset-65			Dataset-360			Dataset-693			Dataset-2230		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Best	99.63	98.89	381	99.71	98.87	1905	99.68	98.82	4341	99.78	99.08	14456
Best-RMSD	98.99	97.19	187	99.20	97.56	1100	99.13	97.33	2291	99.37	97.83	8254

RMSD (promedio y desviación estándar)

	Dataset-65	Dataset-360	Dataset-694	Dataset-2230
Best	0.665 ± 0.15	0.641 ± 0.15	0.651 ± 0.16	0.618 ± 0.15
Best-RMSD	0.534 ± 0.07	0.529 ± 0.07	0.531 ± 0.07	0.518 ± 0.07

Tabla 9. Máxima precisión alcanzable para cada conjunto de prueba.

Precisión total y número de colisiones

	Dataset-65			Dataset-360			Dataset-693			Dataset-2230		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Best	**99.63**	**98.89**	381	**99.71**	**98.87**	1905	**99.68**	**98.82**	4341	**99.78**	**99.08**	14456
Best (CIS-RR)	99.06	97.23	522	99.03	97.05	2695	99.02	97.20	5813	99.13	97.43	21321
Best (RASP)	99.22	97.91	410	99.26	97.79	2254	99.21	97.79	5023	99.30	98.03	17773
Best+	98.34	96.60	**11**	98.18	96.21	**48**	98.13	96.08	**100**	98.29	96.50	**405**
Best+ (CIS-RR)	97.82	94.95	22	97.63	94.67	107	97.65	94.78	334	97.73	95.03	1375
Best+ (RASP)	98.01	95.75	20	97.90	95.40	129	97.87	95.36	329	97.96	95.65	1378

RMSD (promedio y desviación estándar)

	Dataset-65	Dataset-360	Dataset-694	Dataset-2230
Best	**0.665** ± **0.15**	**0.641** ± **0.15**	**0.651** ± **0.16**	**0.618** ± **0.15**
Best (CISRR)	0.773 ± 0.18	0.764 ± 0.19	0.767 ± 0.18	0.745 ± 0.19
Best (RASP)	0.707 ± 0.17	0.708 ± 0.18	0.712 ± 0.18	0.686 ± 0.18
Best+	0.825 ± 0.23	0.835 ± 0.24	0.855 ± 0.26	0.816 ± 0.25
Best+ (CISRR)	0.915 ± 0.23	0.913 ± 0.24	0.919 ± 0.25	0.895 ± 0.25
Best+ (RASP)	0.857 ± 0.24	0.870 ± 0.24	0.875 ± 0.25	0.849 ± 0.25

Capítulo 4. Evaluación de funciones de score para el PSCPP

En el capítulo anterior se presentó la metodología y los resultados en cuanto a la evaluación y comparación de los métodos del estado del arte al considerar los contactos simétricos; así como el cálculo de la máxima precisión alcanzable con una biblioteca de rotámeros estándar (Dunbrack y Cohen, 1997). En este capítulo se describe el método de evaluación de funciones de score para el PSCPP propuesto en este trabajo, el cual se basa en un algoritmo de búsqueda local; así como los resultados de su aplicación en las funciones de score de dos métodos del estado del arte. La mayor parte del contenido sobre estos temas se basa en un artículo de revista producto de este trabajo de investigación (Colbes *et al.*, 2016). Finalmente, en este capítulo también se propone la utilización de algoritmos basados en la búsqueda local como métodos estándar de búsqueda para el PSCPP.

4.1. Funciones de score implementadas

Para todas las pruebas realizadas que se presentan en este capítulo (y también en el siguiente) se emplearon las funciones de score de CIS-RR (Cao *et al.*, 2011) y RASP (Miao *et al.*, 2011). Éstas se eligieron debido a que existe información detallada sobre las mismas que facilita su implementación. Además, aunque las definiciones de los términos empleados en los métodos del estado del arte suelen ser distintas entre sí, las interacciones consideradas son esencialmente las mismas.

4.1.1. CIS-RR

La función de score es una versión adaptada de la propuesta para SCWRL3 (Canutescu *et al.*, 2003), el cual consiste de los siguientes términos:

$$E = E_{VdW} + k_{Rot}E_{Rot} + E_{SS} \tag{6}$$

Donde E_{VdW} es un potencial de Van der Waals empírico, E_{Rot} es el término correspondiente a las preferencias de los rotámeros en la biblioteca (llamado también *término*

de rotámeros) y k_{Rot} su peso asociado; mientras que E_{SS} considera los enlaces disulfuro entre cisteínas.

4.1.1.1. Interacciones de Van der Waals

La interacción $E_{VdW}(i,j)$ entre dos átomos i y j se define de la siguiente manera:

$$E_{VdW}(i,j) = \begin{cases} k_{rep}(1 - r_{ij}/R_{ij}) & \text{para } r_{ij} \leq R_{ij} \\ k_{att}\left[(r_{ij}/R_{ij})^2 - 3(r_{ij}/R_{ij}) + 2\right] & \text{para } R_{ij} < r_{ij} < 2R_{ij} \\ 0 & \text{para } r_{ij} \geq 2R_{ij} \end{cases} \tag{7}$$

Donde r_{ij} es la distancia entre los dos átomos y R_{ij} es la suma de sus radios de Van der Waals[1] (todos en Åmströngs), los cuales se obtuvieron de C. Chothia (1976) con una reducción del 5 % determinada por los autores de CIS-RR. Estos valores se encuentran en el Anexo A.3.

4.1.1.2. Preferencia del rotámero

Este término expresa la preferencia de la conformación de la cadena lateral contenida en un rotámero, y esta preferencia está directamente relacionada a la probabilidad del rotámero en la biblioteca empleada. Así, para el rotámero r_m del residuo m:

$$E_{Rot}(r_m) = -\gamma \log\left(\frac{p(r_m|\phi,\psi)}{p(\text{máx}\,|\phi,\psi)}\right) \tag{8}$$

Donde $p(r_m|\phi,\psi)$ es la probabilidad del rotámero para los ángulos de torsión de la cadena principal ϕ y ψ dados. Este valor se normaliza a la máxima probabilidad $p(\text{máx}\,|\phi,\psi)$ para el residuo m en los mismos valores de ϕ y ψ (Canutescu *et al.*, 2003). Dado que las cadenas laterales de los residuos aromáticos son más propensos a las colisiones, sus valores de E_{Rot} podrían estar subestimados comparados con sus valores de E_{VdW}.

[1] La definición se encuentra en la Subsección 3.6.4.

Por ello se introduce el factor γ, teniéndose que:

$$\gamma = \begin{cases} 2 & \text{si } m \in \{\text{HIS, PHE, TYR, TRP}\} \\ 0 & \text{si } m \in \{\text{GLY, ALA}\} \\ 1 & \text{en caso contrario} \end{cases} \quad (9)$$

4.1.1.3. Enlaces disulfuro

Se emplea la fórmula $E_{SS} = 0.1S - 4.0$ para evaluar todos los pares de cisteínas en la proteína considerada, los cuales pueden formar enlaces disulfuro. S es una puntuación empírica de enlaces disulfuro, considerando dos cisteínas i y j, definida en SCWRL3 (Canutescu *et al.*, 2003) como:

$$S = \frac{|d-2|}{0.05} + \frac{|A_i - 104|}{5} + \frac{|A_j - 104|}{5} + \frac{\text{mín}\{||\chi_2|-80|, ||\chi_2|-180|\}}{10} + \frac{\text{mín}\{||\chi_4|-80|, ||\chi_4|-180|\}}{10} + \frac{||\chi_3|-90|}{20} + \frac{E_{Rot}(r_i) + E_{Rot}(r_j)}{2} \quad (10)$$

Donde d es la distancia entre los átomos $S_{\gamma i}$ y $S_{\gamma j}$, A_i es el ángulo entre los enlaces en $C_{\beta i} - S_{\gamma i} - S_{\gamma j}$, A_j es el ángulo entre los enlaces en $S_{\gamma i} - S_{\gamma j} - C_{\beta j}$, χ_2 es el ángulo de torsión en $C_{\alpha i} - C_{\beta i} - S_{\gamma i} - S_{\gamma j}$, χ_4 es el ángulo de torsión en $S_{\gamma i} - S_{\gamma j} - C_{\beta j} - C_{\alpha j}$, χ_3 es el ángulo de torsión en $C_{\beta i} - S_{\gamma i} - S_{\gamma j} - C_{\beta j}$, y r_i y r_j son los rotámeros para las cisteínas consideradas. Se considera que se forma un enlace disulfuro cuando $E_{SS} < 0$, y su valor se añade al score total de la proteína.

Finalmente, los parámetros (k_{rot} = 0.35, k_{rep} = 5.882, k_{att} = 0.35) de la función de score se obtuvieron a partir del entrenamiento con un conjunto de 55 proteínas, empleando un criterio que se explicará más adelante con mayor detalle.

4.1.2. RASP

La función de score de RASP tiene la siguiente forma:

$$E = E_{VdW} + E_{Rot} + E_{SS} + E_{O,H} \quad (11)$$

En relación a CIS-RR, en este caso se añade la consideración de los enlaces de hidrógeno $E_{O,H}$; aunque la implementación de los otros tres términos es diferente a la de CIS-RR.

4.1.2.1. Interacciones de Van der Waals

El potencial de Van der Waals E_{VdW} es una adaptación del empleado en OPUS-Rota (Lu *et al.*, 2008a), formulado de la siguiente manera para dos átomos i y j:

$$E_{VdW}(i,j) = \begin{cases} 50e_{ij} & \text{si } d' < 0.465 \\ e_{ij}(80 - 64.5d') & \text{para } 0.465 \leq d' < 0.75 \\ 1.63e_{ij}\left[\left(\frac{1}{d'}\right)^{12} - 2\left(\frac{1}{d'}\right)^{6}\right] & \text{para } 0.75 \leq d' < 0.8929 \\ 0.99e_{ij}\left[\left(\frac{1}{d'}\right)^{12} - 2\left(\frac{1}{d'}\right)^{6}\right] & \text{para } 0.8929 \leq d' < 2.3 \\ 0 & \text{para } 2.3 < d' \end{cases} \quad (12)$$

Donde $e_{ij} = \sqrt{e_i e_j}$ es la media geométrica de las profundidades[2] e_i y e_j obtenidas en los parámetros de charmm19 (Brooks *et al.*, 2009). El valor $d' = d_{ij}/R_{ij}$ expresa la relación entre la distancia d_{ij} entre los átomos y la suma R_{ij} de sus radios de Van der Waals. Las constantes 1.63 y 0.99 expresan la diferencia entre los efectos repulsivos y atractivos. El término repulsivo está limitado a un máximo de $50e_{ij}$ para mitigar el hecho de emplear rotámeros fijos. Todos los parámetros necesarios para implementar éste y otros términos se encuentran en el Anexo A.3.

4.1.2.2. Preferencia del rotámero

Este término toma la forma empleada en SCWRL4 (Krivov *et al.*, 2009) y difiere ligeramente de la empleada en CIS-RR:

$$E_{Rot}(r_m) = -w_{aa} \log\left(\frac{p(r_m|\phi,\psi)}{p(\text{máx}\,|\phi,\psi)}\right) \quad (13)$$

La diferencia principal con respecto a CIS-RR está en el uso de un factor w_{aa} que es distinto para cada tipo de aminoácido "aa" considerado (los valores se encuentran en

[2] La definición se encuentra en la Subsección 3.3.2.1.

el Anexo A.3).

4.1.2.3. Enlaces disulfuro

Este término es una versión simplificada del que se emplea en SCWRL3 (Canutescu *et al.*, 2003):

$$S = 6\left(|d-2.06| + \frac{|A_i-105|+|A_j-105|}{100} + \frac{||\chi_3|-90|}{140}\right) - 11.4 \qquad (14)$$

Donde d es la distancia del enlace $S_\gamma - S_\gamma$, A_i y A_j son los dos ángulos entre enlaces $S_\gamma - S_\gamma - C_\beta$ y χ_3 es el ángulo de torsión $C_\beta - S_\gamma - S_\gamma - C_\beta$. La energía de un enlace disulfuro estándar es de 11.4 kcal/mol (Miao *et al.*, 2011).

4.1.2.4. Enlaces de hidrógeno

Este término considera únicamente enlaces de hidrógeno entre hidroxilos (-OH) y carboxilos (-COOH), y está formulado de la siguiente manera:

$$E_{O,H} = -1.8\sqrt{\frac{(\cos(\alpha-111.5)-\cos 37)(\cos(\beta-120)-\cos 47)}{(1-\cos 37)(1-\cos 47)}} \qquad (15)$$

Donde α es el ángulo entre los segmentos $B_1 - D_1$ y $D_1 - O_2$, y β es el ángulo entre los segmentos $D_1 - O_2$ y $O_2 - C_2$. Aquí D_1 es el donador de hidrógeno (átomo de oxígeno del hidroxilo), B_1 es la base del donador, O_2 es el aceptor de hidrógeno (átomo de oxígeno del carboxilo) y C_2 es el átomo de carbono del carboxilo. Este término sólo se calcula si la distancia entre D_1 y O_2 es menor a 3.2 Å. Dado que no se necesitan coordenadas explícitas de átomos de hidrógeno en este término, su cálculo es relativamente muy rápido (Miao *et al.*, 2011).

4.1.3. Algunas consideraciones en el cálculo de interacciones

Como se mencionó en la Subsección 3.3.2, los términos de las funciones de score para el PSCPP usualmente se limitan a considerar interacciones internas de un residuo e interacciones entre pares de residuos; y puede verse que todos los términos

en CIS-RR y RASP cumplen con esta característica. Ésto permite una implementación más rápida de los métodos para el PSCPP: las interacciones se pueden pre-calcular considerando todas las posibilidades dadas por los rotámeros para cada residuo; para así almacenarlas en estructuras de datos que permitan una evaluación más rápida de soluciones candidatas durante el proceso de búsqueda. En el caso de las interacciones internas en un residuo, los valores calculados se almacenan en un arreglo de dos dimensiones A, donde $A(i,k)$ expresa el valor de las interacciones en el residuo i teniendo el rotámero k. Para las interacciones entre pares de residuos se emplea un arreglo de cuatro dimensiones B, donde $B(i,k,j,l)$ expresa el valor de las interacciones entre el residuo i con la conformación determinada por el rotámero k, y el residuo j con la conformación dada por el rotámero l.

Si una solución candidata se expresa mediante un arreglo S que indica para cada residuo (en el orden dado por la secuencia de N residuos de la proteína) la posición del rotámero en la biblioteca empleada, entonces el valor en la función de score para esta solución está dado por:

$$Score_S = \sum_{i=1}^{N} A(i, S(i)) + \sum_{i=1}^{N} \sum_{j=1}^{N} B(i, S(i), j, S(j)) \tag{16}$$

Las interacciones entre los átomos son nulas o al menos despreciables a partir de cierta distancia. Por esta razón se considera algún criterio para descartar ciertas interacciones entre pares de residuos, para así acelerar el cálculo del arreglo B. El criterio empleado en el presente trabajo se presenta en el Anexo A.4.

4.2. Método de búsqueda local para la evaluación de funciones de score para el PSCPP

Para la predicción de estructuras de proteínas es común usar los llamados *conjuntos de "señuelos"* (*decoy sets* en inglés) para la evaluación de funciones de score. Los señuelos son estructuras similares a la determinada experimentalmente para una proteína, creadas artificialmente mediante diversas técnicas; y son usualmente empleadas para el diseño, entrenamiento y evaluación de funciones de score (Park y Levitt, 1996; Handl *et al.*, 2009). Un conjunto de señuelos generalmente incluye la estructura

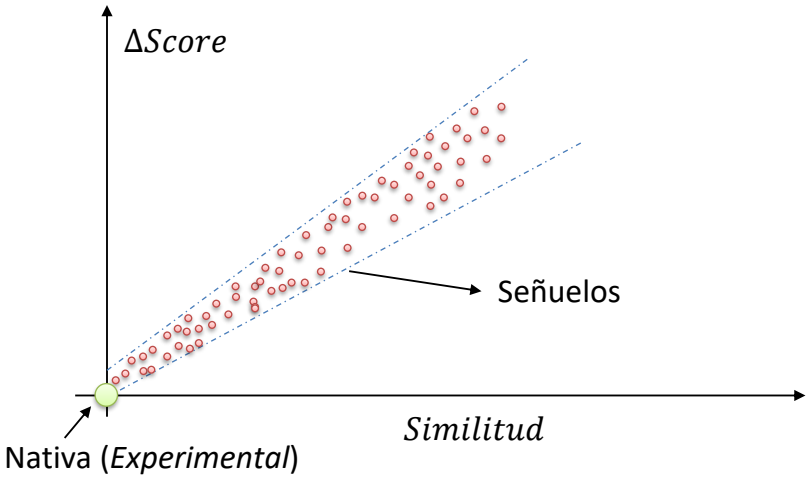

Figura 26. Correlación entre la diferencia de scores y la similitud de un señuelo con la conformación experimental (que representa a la estructura nativa de la proteína). Cuanto más se aproxime la estructura de un señuelo a la experimental, menor será la diferencia de scores.

experimental, y una función de score ideal debería reconocer siempre esta estructura entre los señuelos. También es común el análisis de correlación entre los valores de la función de score y la similitud de los señuelos con la estructura experimental. En este caso, el objetivo es verificar si una función de score es capaz de clasificar señuelos de manera confiable y así proporcionar una guía precisa hacia la estructura experimental (bajo la suposición de que es la nativa) a través del espacio de búsqueda (Handl *et al.*, 2009). Como se muestra en la Figura 26, la idea central es que cuanto más similar sea un señuelo a la estructura experimental, menor será la diferencia (no-negativa) entre sus valores de energía/score. Existen diferentes métricas de similitud empleadas, siendo el RMSD un indicador usual pero con ciertas limitaciones que motivaron la aparición de otras métricas complementarias (Zhang y Skolnick, 2004).

Con respecto al PSCPP, la evaluación usual de cualquier método para este problema involucra a sus tres componentes principales; por lo que se necesita “aislar” de alguna forma la función de score para su evaluación. En el presente trabajo de tesis se propone usar un método de búsqueda local (LS por sus siglas en inglés) para la evaluación de funciones de score para el PSCPP, el cual extiende y mejora otro método conocido como *búsqueda de conformaciones para un solo residuo* (SRCS por sus siglas en inglés) (Petrella *et al.*, 1998; Xiang y Honig, 2001; Liang y Grishin, 2002). En el SRCS, la función de score se minimiza para un solo residuo a la vez, dejando a los demás residuos en sus posiciones en la estructura experimental. Al final de la

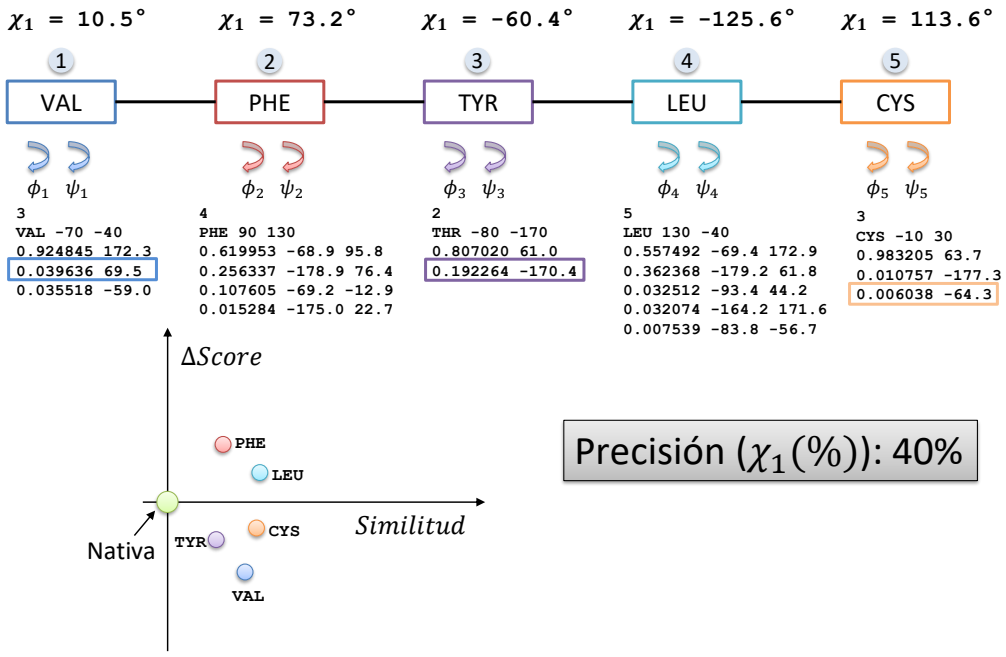

Figura 27. Ejemplo de aplicación de la búsqueda de conformaciones para un solo residuo, en el cual se evalúa la calidad de la función de score mediante la precisión de χ_1(%). Los ángulos de torsión de χ_1 para la conformación de la estructura experimental se encuentran en la parte superior de cada residuo, y en la parte inferior se encuentran los rotámeros correspondientes en la biblioteca. Durante el proceso de evaluación, para cada residuo se cambia temporalmente la configuración experimental por cada uno de los rotámeros, buscando minimizar la función de score (los demás residuos quedan en la configuración experimental). Si existen rotámeros que logran un menor valor en la función evaluada, entonces se determina si el ángulo de torsión del rotámero con menor valor de score es correcto (de acuerdo al criterio de los 40°). La calidad de la función evaluada es entonces el porcentaje de residuos para los cuales el valor de χ_1 es correcto. En el ejemplo se produjeron cambios con ángulos de torsión incorrectos para VAL, TYR y CYS; por lo que la precisión es del 40 %. Este método puede verse también como la evaluación de señuelos que resultan de la modificación de la estructura experimental en un solo residuo.

minimización para un residuo en particular, se determina la precisión de la nueva conformación (según lo establecido en la Subsección 3.6.1). Luego este residuo retorna a su configuración en la estructura experimental y se pasa al siguiente; repitiéndose este proceso hasta llegar al último residuo de la proteína. La precisión de la función de score se mide por el porcentaje de residuos con la conformación correcta de acuerdo a la métrica considerada (χ_1(%) y χ_{1+2}(%), por poner ejemplos). Un ejemplo de su aplicación se muestra en la Figura 27.

Por otro lado, el método LS que se propone aquí consiste en aplicar un algoritmo de búsqueda local, empleando la estructura experimental como entrada. Los primeros intentos de aplicar el método LS se hicieron en los trabajos de Lezcano (2012) y Rodríguez (2014), los cuales consideraron sólo algunos términos de las funciones de score de métodos del estado del arte y se centraron en una estructura *similar* a la

experimental como entrada. El Algoritmo 2 muestra el proceso de búsqueda local: en una iteración se visita cada residuo, y entre las posibilidades (conformación actual + rotámeros correspondientes al residuo) se elige la de mínimo valor en la función de score. La diferencia principal con el SRCS es que al pasar al siguiente residuo ya no se vuelve a la conformación experimental en el residuo actual, sino en la conformación que logra el menor valor en la función de score. Este proceso continúa hasta que en una iteración no se produzcan cambios en algún residuo, lo cual indica que se alcanzó un mínimo local en la función de score.

Algoritmo 2 Algoritmo de búsqueda local.

Entrada: Estructura S_{in}
Salida: Estructura S_{LS}
1: $S_{LS} = S_{in}$
2: $min_{score} = score(S_{LS})$
3: **repetir**
4: $fin = Verdadero$
5: **para** cada residuo i **hacer**
6: **para** cada rotámero r de i **hacer**
7: Asignar temporalmente r a i, obteniendo S_r
8: sc = score(S_r)
9: **si** $sc < min_{score}$ **entonces**
10: $min_{score} = sc$
11: $S_{LS} = S_r$
12: $fin = Falso$
13: **fin si**
14: **fin para**
15: **fin para**
16: **hasta que** $fin == Verdadero$
17: **devolver** S_{LS}

La idea central de aplicar este algoritmo de búsqueda local a la estructura experimental es que la misma debe ser al menos un mínimo local para una buena función de score, por lo que el algoritmo no debería ser capaz de encontrar soluciones vecinas con mejor score que el de la estructura experimental.

El método LS que se propone en este trabajo comparte ciertas características con el análisis de correlación (para la predicción de estructuras de proteínas) y con el SRSC (para el PSCPP), debido a que:

- Considera todos los posibles rotámeros para un residuo a la vez, dejando fijos a los demás residuos.

- Los cambios en la estructura experimental corresponden a señuelos cercanos a él: muchos cambios en la estructura experimental indican valores de score sucesivamente menores a lo largo del camino, por lo que en ese caso la correlación entre los valores de score y la similitud no es buena para la función de score evaluada.

A pesar de que el método LS y el método SRCS (Petrella *et al.*, 1998) aparentan ser similares, la diferencia principal entre ellos se ilustra en la Figura 28. Todos los candidatos examinados durante la ejecución del SRCS se podrían considerar como señuelos que difieren de la estructura experimental únicamente en un rotámero de un solo residuo. Por lo tanto, el indicador de desempeño de la función de score da una cota inferior en el número de estos señuelos con menor score que la estructura experimental (como puede verse en la parte inferior de la Figura 27). Como el conjunto de señuelos considerado por el SRCS es restringido, sólo describe el comportamiento de la función de score en una vecindad muy cercana a la estructura experimental. Un caso extremo se ilustra en la Figura 29, para el cual la función de score tiene un bajo desempeño de acuerdo al SRCS. Sin embargo, la función de score puede considerarse como buena, ya que las estructuras con menores valores de score que el de la estructura experimental están cercanas a ella. Otra limitación del SRSC es que no provee información acerca del comportamiento de la función de score para estructuras que difieren de la experimental en más de un rotámero.

Por el contrario, el método LS proporciona un escenario distinto: cuando se encuentra un señuelo con menor score al de la estructura experimental, luego se evalúa la vecindad de ese señuelo. Y como este proceso es iterativo, potencialmente se puede alcanzar señuelos lejanos a la estructura experimental. El número de cambios de rotámeros es de hecho el número de señuelos sucesivos con menor score encontrados a lo largo del proceso, por lo que puede ser un indicador de la calidad de la función de score. Sin embargo, el número de cambios no es necesariamente proporcional al número de residuos con conformaciones distintas entre la estructura experimental y la que se obtiene al final de la búsqueda local; por lo tanto, una métrica más importante es la variación de la precisión al final de la búsqueda local. Una alta variación de precisión indica que se alcanzó la estructura resultante a partir de la experimental mediante señuelos sucesivos con menores valores de score, por lo que es improbable

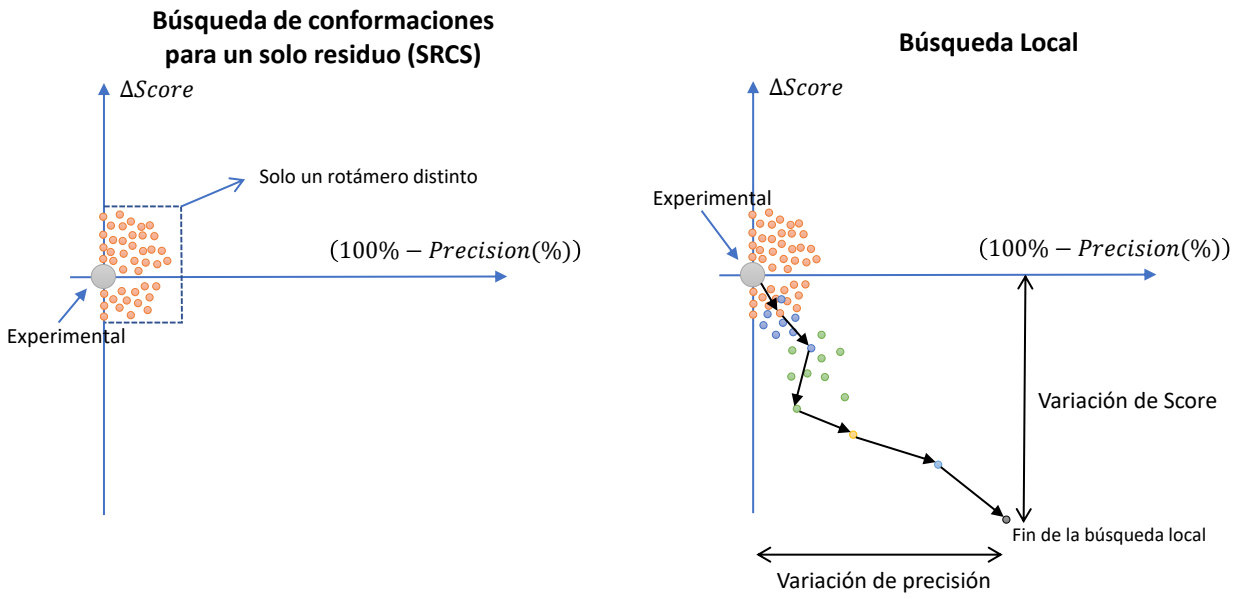

Figura 28. Diferencia entre el método SRSC y el método LS. Δ*Score* es la diferencia de score entre un señuelo y la estructura experimental de referencia.

que un método de búsqueda pudiese llegar a la estructura experimental mediante la función de score evaluada.

La Figura 30 muestra la aplicación del método de búsqueda local para la estructura experimental 1MUW de la enzima *xilosa isomerasa*, usando la función de score del RASP. Al final de la búsqueda local se tiene que $\Delta S = -200$ y una pérdida de precisión de aproximadamente 12 %. Puede verse que el SCRS únicamente evalúa la pequeña región a la izquierda (indicada con un círculo), que corresponde a los señuelos con a lo más un rotámero incorrecto. Sin embargo, el LS también da información acerca de la distribución de los valores de score para señuelos que, en cuanto a precisión, están más alejados de la estructura experimental.

Como el método LS proporciona una información más útil acerca del desempeño de una función de score y no está restringido a evaluar únicamente a los señuelos muy cercanos a la estructura experimental, este método se propone como una mejora del método SRCS (Petrella *et al.*, 1998). En el ejemplo representado por la Figura 29, el método LS terminaría en una estructura cercana a la experimental, por lo que el desempeño de la función de score sería casi ideal.

En el presente trabajo se implementaron las funciones de score de CIS-RR (Cao *et al.*, 2011) y RASP (Miao *et al.*, 2011), para evaluarlas mediante el método LS. Las

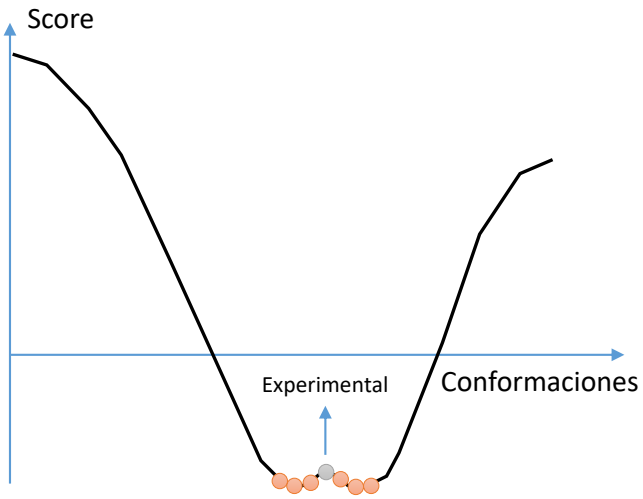

Figura 29. Ejemplo de un caso extremo para una función de score para ilustrar la diferencia principal entre el SRCS y el LS. En este ejemplo, el SRCS indicará un muy mal desempeño de la función de score, ya que los señuelos cercanos a la estructura experimental tienen un menor score. Sin embargo, éste no será el caso para el LS, ya que el mismo terminará en un óptimo local muy cercano a la estructura experimental.

Tablas 10 y 11 muestran los resultados de la evaluación para estas dos funciones de score, considerando cinco tipos de estructuras de entrada para el algoritmo de búsqueda local:

- *Experimental*: para cada residuo de la estructura, su conformación en la estructura experimental se agrega a su lista de posibles rotámeros a explorar. Este *rotámero experimental* se asigna inicialmente al residuo considerado, y su probabilidad es la máxima entre el conjunto de rotámeros correctos para dicho residuo.
- *Best* y *Best+*: como se detalló en la Sección 3.8.
- *MostProb*: cada residuo de la estructura inicial se genera empleando el rotámero con la mayor probabilidad en la biblioteca. Para los conjuntos de prueba en este trabajo, los valores iniciales de χ_1(%) y χ_{1+2}(%) son aproximadamente 73 % y 58 %, respectivamente.
- *Random*: cada residuo de la estructura inicial se genera seleccionando al azar un rotámero de la biblioteca. Para los conjuntos de prueba en este trabajo, los valores iniciales de χ_1(%) y χ_{1+2}(%) son aproximadamente 34 % y 16 %, respectivamente.

Los motivos detrás de la selección de estas estructuras iniciales son los siguientes:

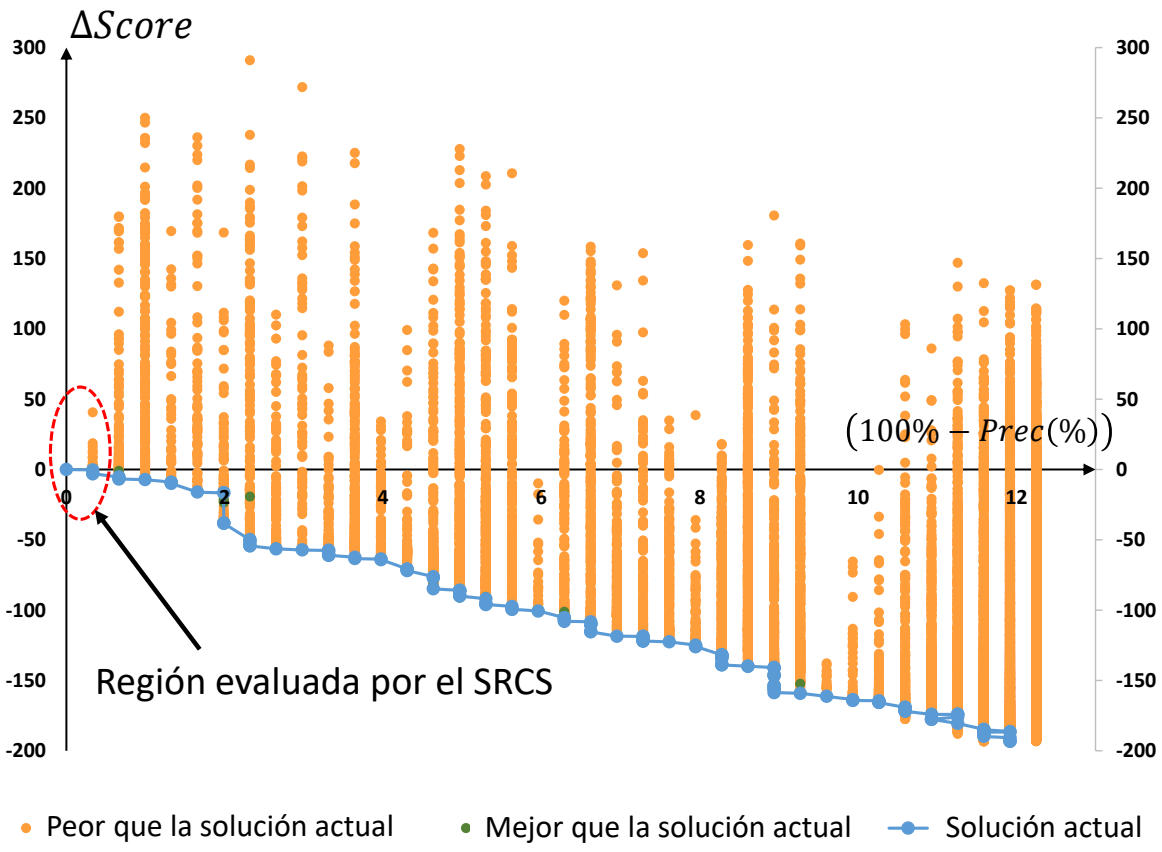

Figura 30. Ejemplo de aplicación del método LS para la función de score del RASP (Miao *et al.*, 2011), empleando la estructura experimental 1MUW de la enzima *xilosa isomerasa*. Los puntos unidos por una línea representan las soluciones intermedias adoptadas durante el proceso de la búsqueda local, y los puntos de color naranja representan a los señuelos con un valor más alto de score que la mejor solución encontrada hasta ese momento por el algoritmo LS (al momento de evaluar el cambio de un rotámero).

- Las estructuras *Experimental*, *Best* y *Best+* tienen información sobre la estructura experimental de la proteína, por lo que se usan para evaluar la función de score mediante el método de búsqueda local.

- Las estructuras *MostProb* y *Random* no emplean información de la estructura experimental, y considerando la simplicidad del algoritmo de búsqueda local, los resultados de precisión obtenidos pueden considerarse como una cota inferior para una función de score dada. Ésto a su vez podría dar una idea de la contribución de los métodos de búsqueda a los resultados en cuanto a precisión para el PSCPP.

Tabla 10. Resultados para la función de score de CIS-RR, empleando diferentes estructuras para el algoritmo de búsqueda local.

Precisión total y número de colisiones

	Dataset-65			Dataset-360			Dataset-693			Dataset-2230		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Experimental	91.72	86.12	3	90.73	84.63	5	91.5	85.64	15	91.19	85.3	44
Best	88.56	79.66	36	87.33	77.74	149	88.2	78.64	336	87.97	78.51	1218
Best+	88.58	79.66	22	87.23	77.62	108	88.03	78.42	239	87.83	78.28	956
MostProb	85.32	74.5	82	84.38	73.38	385	85.01	73.7	867	84.62	73.53	3269
Random	83.93	72.98	128	83.12	71.66	533	83.85	72.26	1197	83.81	72.35	3953

Promedio de los valores de score y número de estructuras finales con menor score que la experimental (#Lower)

	DS-65 ($E_{exp} = -0.925$)			DS-360 ($E_{exp} = -0.899$)			DS-693 ($E_{exp} = -0.979$)			DS-2230 ($E_{exp} = -0.844$)		
	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower
Experimental	-0.925	-1.145	65	-0.899	-1.223	360	-0.979	-1.207	692	-0.844	-1.190	2230
Best	-0.235	-0.933	41	-0.124	-0.999	300	-0.221	-0.981	471	-0.123	-0.978	1716
Best+	-0.309	-0.935	41	-0.156	-0.998	299	-0.273	-0.981	472	-0.173	-0.978	1711
MostProb	2.809	-0.867	33	3.131	-0.931	251	3.153	-0.910	342	2.933	-0.911	1390
Random	8.794	-0.814	24	9.798	-0.883	213	9.354	-0.864	276	9.660	-0.878	1273

E_{exp} es el score de la estructura experimental, E_{start} es el score de la estructura inicial, and E_{end} es el score de la estructura al final de la búsqueda local.

Promedio del número de cambios de rotámeros y RMSD promedio

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios
Experimental	1.118 ± 0.22	125.3	1.231 ± 0.24	144.5	1.145 ± 0.23	128.5	1.218 ± 0.28	157.0
Best	1.426 ± 0.21	93.3	1.523 ± 0.23	112.3	1.459 ± 0.23	97.9	1.513 ± 0.27	119.3
Best+	1.423 ± 0.21	92.4	1.534 ± 0.23	112.0	1.474 ± 0.24	97.5	1.524 ± 0.27	119.0
MostProb	1.682 ± 0.21	105.2	1.740 ± 0.28	116.7	1.699 ± 0.27	114.7	1.745 ± 0.30	127.1
Random	1.754 ± 0.29	227.8	1.828 ± 0.28	254.5	1.785 ± 0.28	242.7	1.807 ± 0.31	283.6
Random(Start)	3.459 ± 0.18	-	3.454 ± 0.22	-	3.462 ± 0.21	-	3.457 ± 0.26	-

La última fila muestra el valor promedio de RMSD para las estructuras Random iniciales.

Tabla 11. Resultados para la función de score de RASP, empleando diferentes estructuras para el algoritmo de búsqueda local.

Precisión total y número de colisiones

	Dataset-65			Dataset-360			Dataset-693			Dataset-2230		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Experimental	91.92	86.81	5	91.14	85.57	25	91.93	86.52	47	91.67	86.39	166
Best	88.96	80.27	53	87.64	78.29	225	88.47	79.15	478	88.23	79.1	1725
Best+	88.84	80.06	37	87.53	78.13	196	88.31	78.96	381	88.08	78.86	1434
MostProb	85.76	74.8	123	84.31	73.37	522	85.11	73.78	1242	84.8	73.82	4216
Random	84.8	73.42	159	83.26	71.72	682	84.22	72.62	1325	83.86	72.57	4893

Promedio de los valores de score y número de estructuras finales con menor score que la experimental (#Lower)

	DS-65 ($E_{exp} = -2.231$)			DS-360 ($E_{exp} = -2.026$)			DS-693 ($E_{exp} = -2.726$)			DS-2230 ($E_{exp} = -2.132$)		
	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower
Experimental	-2.231	-3.007	65	-2.026	-3.185	360	-2.726	-3.530	691	-2.132	-3.301	2229
Best	0.083	-2.356	41	0.662	-2.487	308	-0.119	-2.828	495	0.366	-2.640	1775
Best+	-0.130	-2.357	42	0.568	-2.481	308	-0.270	-2.827	496	0.230	-2.637	1773
MostProb	8.475	-2.183	33	9.472	-2.317	275	8.997	-2.628	360	8.629	-2.460	1470
Random	25.655	-2.099	28	28.803	-2.177	236	26.858	-2.538	323	28.338	-2.371	1344

E_{exp} es el score de la estructura experimental, E_{start} es el score de la estructura inicial, and E_{end} es el score de la estructura al final de la búsqueda local.

Promedio del número de cambios de rotámeros y RMSD promedio

	Dataset-65		Dataset-360		Dataset-693		Dataset-2230	
	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios
Experimental	1.083 ± 0.21	121.4	1.186 ± 0.25	139.5	1.106 ± 0.23	124.3	1.179 ± 0.29	150.9
Best	1.420 ± 0.22	92.6	1.497 ± 0.24	111.9	1.438 ± 0.23	97.8	1.498 ± 0.28	119.3
Best+	1.425 ± 0.21	91.7	1.508 ± 0.24	111.8	1.455 ± 0.24	97.7	1.511 ± 0.28	119.1
MostProb	1.658 ± 0.21	112.0	1.731 ± 0.29	122.4	1.698 ± 0.27	119.3	1.733 ± 0.31	133.5
Random	1.701 ± 0.24	227.3	1.799 ± 0.27	254.9	1.758 ± 0.27	242.2	1.795 ± 0.30	282.6
Random(Start)	3.448 ± 0.26	-	3.474 ± 0.21	-	3.457 ± 0.2	-	3.458 ± 0.25	-

La última fila muestra el valor promedio de RMSD para las estructuras Random iniciales.

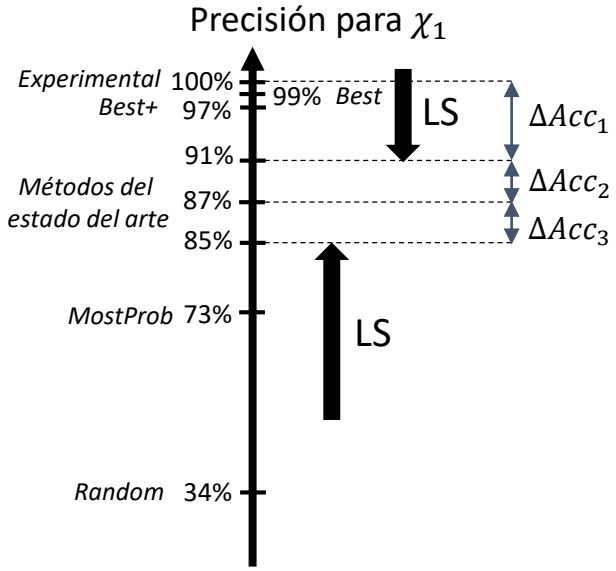

Figura 31. Precisiones para χ_1(%) de las estructuras iniciales y las obtenidas al final de la búsqueda local. ΔAcc_1 indica la pérdida de precisión al final de la búsqueda local, empezando con la estructura experimental. ΔAcc_2 indica la diferencia entre el promedio de las precisiones de los métodos del estado del arte evaluados en la Subsección 3.7.8, y las obtenidas al final de la búsqueda local (empezando con la estructura experimental). ΔAcc_3 es un aproximado de la ganancia de precisión lograda por mejores métodos de búsqueda que el algoritmo LS. Para una buena función de score debe cumplirse que: $\Delta Acc_1 \ll \Delta Acc_2$.

Estas ideas se ilustran en la Figura 31. Los resultados para χ_1(%) en las Tablas 10 y 11 revelan que la pérdida de precisión al final de la búsqueda local, comenzada con la estructura experimental (ΔAcc_1), es mayor al 8 %. Esto indica que estas estructuras finales están mucho más cerca a los resultados de los métodos del estado del arte que a los valores ideales.

Como el score de una estructura está relacionada a su cantidad de residuos, los valores promedio para cada conjunto de prueba pueden estar influenciados por los valores de score de proteínas grandes. Se ha observado que esta relación es aproximadamente lineal para ambas funciones de score evaluadas; por lo tanto, los valores presentados en las Tablas 10 y 11 están normalizados. Puede verse que, para ambas funciones de score, a pesar de que las estructuras *Best* tienen un valor más alto de precisión, existe una diferencia de score significativa con respecto a la estructura experimental (aunque son valores promedio, esta diferencia fue observada en casi todas las proteínas de cada conjunto de prueba). Además, los valores de score al final de la búsqueda local se encuentran mucho más cerca a los correspondientes a las estructuras experimentales. Inicialmente se pensó que esta diferencia se debe a la gran

cantidad de colisiones en las estructuras *Best*; sin embargo, los scores iniciales de las estructuras *Best+* no mostraron mejoras significativas.

Dado el número de casos en donde el score al final de la búsqueda local es menor al score de la estructura experimental (incluso empezando desde estructuras *Random*, que son estructuras iniciales obtenidas al azar), no se puede garantizar que se obtendrá la estructura experimental al minimizar estas funciones de score. En consecuencia, por más bueno que sea el método de búsqueda utilizado, no se podrán obtener resultados cercanos a los ideales. Esto resulta más evidente para las estructuras experimentales, donde para casi todos los casos existen estructuras con menores valores de score al final de la búsqueda local. Nótese que el número de veces que se mejora el score de la estructura experimental es mayor cuando se empieza la búsqueda desde las estructuras experimentales. Esto se debe a la forma de la gráfica (*landscape*) inducida por la función de score, sumado al método de búsqueda local: cuanto más lejos de la estructura experimental comience la búsqueda, mayor es la probabilidad de estancarse en un mínimo local con peor calidad que la correspondiente a la estructura experimental.

En la Figura 31 puede verse que existe una brecha de aproximadamente 2 % para el valor de χ_1(%) entre los métodos del estado del arte y un algoritmo simple de búsqueda local, el cual inicia con una estructura cuyos residuos se generan con los rotámeros de mayor probabilidad en la biblioteca. Esta diferencia sugiere que los esfuerzos para mejorar los resultados para el PSCPP deben enfocarse en las funciones de score en lugar de las estrategias de búsqueda, lo cual también fue conjeturado por (Liang y Grishin, 2002).

Se tuvo un particular interés en los rotámeros seleccionados para formar las estructuras al final de la búsqueda local empezando de la experimental, con el fin de determinar si estos scores se pudiesen alcanzar sin los rotámeros de la estructura experimental. La Tabla 12 muestra los resultados de reemplazar los rotámeros experimentales por otros *similares* (bajo el mismo criterio utilizado para la construcción de estructuras *Best*) en la biblioteca, los cuales sugieren que estos scores se obtuvieron únicamente por la presencia de estos rotámeros experimentales.

Tabla 12. Efecto de reemplazar los *rotámeros experimentales* al final de la búsqueda local aplicada a estructuras experimentales.

CIS-RR

Dataset	$E_{LS_{Exp}}$	$E_{BestLS_{Exp}}$	#Lower
Dataset-65	-1.145	-0.561	0
Dataset-360	-1.223	-0.564	0
Dataset-693	-1.207	-0.559	0
Dataset-2230	-1.190	-0.554	2

RASP

Dataset	$E_{LS_{Exp}}$	$E_{BestLS_{Exp}}$	#Lower
Dataset-65	-3.007	-1.293	0
Dataset-360	-3.185	-1.198	0
Dataset-693	-3.530	-1.473	0
Dataset-2230	-3.301	-1.338	2

- LS_{Exp}: estructura al final de la búsqueda local, comenzando con la estructura experimental.
- $BestLS_{Exp}$: estructura que resulta al reemplazar los *rotámeros experimentales* por los más cercanos en la biblioteca.
- $E_{LS_{Exp}}$: score de LS_{Exp}
- $E_{BestLS_{Exp}}$: score de $BestLS_{Exp}$
- #Lower: número de estructuras con menor score que la experimental después del reemplazo.

4.3. Evaluación de algoritmos basados en la búsqueda local

Hasta este punto, se tiene que la máxima precisión alcanzable usando una biblioteca de rotámeros estándar (Dunbrack y Cohen, 1997) es 97 % para χ_1(%) y 94 % para χ_{1+2}(%); sin embargo, la alcanzada por los métodos del estado del arte es de 87 % para χ_1(%) y 77 % para χ_{1+2}(%). Por lo tanto, existe una brecha significativa de mejora para el PSCPP, y esta biblioteca de rotámeros es suficiente para lograr resultados casi ideales.

De los resultados de la sección anterior se desprende que las funciones de score son las principales responsables de las limitaciones actuales para el PSCPP. Para centrar los esfuerzos hacia el desarrollo de mejores funciones de score (*FS* de aquí en adelante) y poder evaluar el desempeño de una FS candidata directamente, se cree importante establecer un algoritmo de búsqueda estándar para el PSCPP. En la sección anterior también se vió que un método para el PSCPP que consta de: (i) la biblioteca de Dunbrack y Cohen (1997), (ii) la función de score de CIS-RR o RASP, y (iii) un algoritmo de búsqueda local con una estructura inicial *MostProb*[3]; logra una precisión cercana a la de los métodos del estado del arte: 85 % para χ_1(%) y 74 % para χ_{1+2}(%).

El algoritmo de búsqueda local es una mejora iterativa de la solución para un pro-

[3]Ver Sección 4.2.

blema de optimización en un determinado momento, considerando todas las soluciones en su vecindario (Aarts y Lenstra, 2003). En el presente trabajo, la función de vecindario está dada por todas las estructuras con sólo un rotámero de diferencia respecto a la estructura actual. El funcionamiento de este algoritmo se explicó con más detalle en la Sección 4.2.

A pesar de los relativamente buenos resultados obtenidos por la búsqueda local en el PSCPP, el principal problema está en que las fisonomías de las gráficas (o *paisajes*[4]) determinadas por las funciones de score son usualmente multimodales (es decir, existen múltiples mínimos locales), por lo que su desempeño es altamente dependiente de la solución inicial. Para superar este problema, existen variaciones del algoritmo de búsqueda local que realizan movimientos aleatorios en un intento de escapar de un óptimo local. Algunos de estos algoritmos son muy usados en la predicción de estructuras de proteínas (Raman *et al.*, 2009; Xu y Zhang, 2012) y el diseño de proteínas (Dahiyat y Mayo, 1996; Liu y Kuhlman, 2006; Khoury *et al.*, 2014). De hecho, OPUS-Rota (Lu *et al.*, 2008a), CIS-RR (Miao *et al.*, 2011) y RASP (Miao *et al.*, 2011) utilizan una versión modificada del algoritmo de búsqueda local.

Por todo lo anterior, se realizó una comparación cuantitativa de cinco algoritmos basados en la búsqueda local aplicados al PSCPP. Es importante recordar que en la Sección 3.5 se estableció que una solución candidata para el PSCPP está determinada por un arreglo de números enteros, donde el elemento en la posición i corresponde a la ubicación, en la biblioteca, del rotámero asignado al residuo i de la proteína. Los cinco algoritmos de búsqueda seleccionados se detallan a continuación:

- Búsqueda local de múltiples inicios (*Multi-Start Local Search, MSLS*) (Aarts y Lenstra, 2003): consiste de múltiples ejecuciones del algoritmo de búsqueda local con arreglos diferentes, generados de manera aleatoria.
- Búsqueda local iterativa (*Iterated Local Search, ILS*) (Lourenço *et al.*, 2010): cuando se alcanza un mínimo local, se realiza una perturbación de la solución y se aplica nuevamente la búsqueda local a la modificación resultante. Así, existe una secuencia de soluciones generadas por la búsqueda local. En los experimentos del presente trabajo, se acepta un nuevo mínimo local sólo si mejora el mejor

[4]Se emplea el término “landscape” en inglés para referirse a la forma de la gráfica determinada por una función.

mínimo local hasta ese momento. Para la perturbación, cada residuo tiene una probabilidad de 0.2 de cambiar a un rotámero de la biblioteca elegido al azar.

- Recocido simulado (*Simulated Annealing, SA*) (Kirkpatrick *et al.*, 1983): existe una probabilidad positiva de aceptar una solución peor que la actual siguiendo el *criterio de Metropolis* (Metropolis *et al.*, 1953), y su valor depende de los scores de las dos soluciones y un parámetro de *temperatura* decreciente. En una iteración de la implementación en este trabajo del SA, los residuos se visitan exactamente una vez en un orden aleatorio. Para un determinado residuo, se selecciona un rotámero de la biblioteca al azar para reemplazar al actual. La temperatura se reduce gradualmente a su valor mínimo (se emplea un valor de 0.05, al igual que OPUS-Rota). Para el cálculo de la temperatura inicial se emplea el método de Kirkpatrick *et al.* (1983): se comienza con una temperatura alta, y hasta que el porcentaje de aceptación de cambios en una iteración es menor al 60 %, se divide la temperatura actual entre dos.

- Recocido simulado sin rechazo (*Rejectionless Simulated Annealing, SARL*) (Greene y Supowit, 1986): en lugar de seleccionar un posible cambio y evaluarlo con el criterio de Metropolis (Metropolis *et al.*, 1953) como en el caso del SA; se almacena una lista de los efectos de cada cambio, y esta información se emplea para sesgar la selección del cambio. En el presente trabajo se siguió la implementación de OPUS-Rota (Lu *et al.*, 2008a), donde los posibles cambios son los rotámeros de la biblioteca que corresponden a un residuo dado. Para cada posible rotámero i, $1 \leq i \leq M$, se almacena $w_i = \exp(-\Delta S_i/T)$, donde ΔS_i es la variación de score que resultaría al adoptar el rotámero i. El cambio al rotámero i se selecciona con una probabilidad dada por:

$$p_i = \frac{w_i}{\sum_{j=1}^{M} w_j} \tag{17}$$

  El procedimiento de reducción de la temperatura es el mismo que para el SA.

- Recocido simulado de OPUS-Rota (*OPUS-Rota Simulated Annealing, SAOPUS*): esta es la implementación del método de búsqueda presentado por Lu *et al.* (2008a) para OPUS-Rota. Es un algoritmo SARL con casi el mismo procedimiento de reducción de temperatura, donde la temperatura inicial se fija a 2.5.

Para el MSLS y el ILS se utilizaron como entrada arreglos generados de manera

Algoritmo 3 Algoritmo de recocido simulado.

Entrada: Arreglo de rotámeros iniciales $R_{in}, T_{min}, numIter$
Salida: Arreglo de rotámeros finales R_{SA}
1: $T = \text{calcularTemperaturaInicial}(R_{in})$
2: $\Delta T = (T - T_{min})/(numIter - 1)$
3: $R_{SA} = R_{in}$
4: $score_{actual} = score(R_{SA})$
5: **repetir**
6: **para** cada residuo i en orden aleatorio **hacer**
7: Elegir un rotámero r para i de manera aleatoria
8: Asignar temporalmente r a i, obteniendo R_r
9: $sc = \text{score}(R_r)$
10: **si** $sc < score_{actual}$ **entonces**
11: $score_{actual} = sc$
12: $R_{SA} = R_r$
13: **si no**
14: $prob = \exp((score_{actual} - sc)/T)$
15: **si** $\text{Random}(0, 1) < prob$ **entonces**
16: $score_{actual} = sc$
17: $R_{SA} = R_r$
18: **fin si**
19: **fin si**
20: **fin para**
21: $T = T - \Delta T$
22: **hasta que** $T < T_{min}$
23: **devolver** R_{SA}

aleatoria. Para los demás se emplearon arreglos de rotámeros donde cada elemento indica la posición del rotámero de mayor probabilidad en la biblioteca para el residuo correspondiente, lo cual también se realizó en trabajos anteriores (Lu *et al.*, 2008a; Cao *et al.*, 2011; Miao *et al.*, 2011).

No se consideró otros tipos de heurísticas (por ejemplo, el algoritmo de campo medio autoconsistente - SCMF (Koehl y Delarue, 1994)) ni metaheurísticas (por ejemplo, un algoritmo genético (Davis, 1991)) porque son más lentos y no proporcionan un beneficio adicional en la minimización de las funciones de score, como lo había señalado Voigt *et al.* (2000).

Para proporcionar una opción que permita controlar el tiempo de ejecución y para asegurar las mismas condiciones en términos de esfuerzo de los métodos de búsqueda seleccionados, se limitó el número de evaluaciones de la función de score para cada uno de ellos. Este límite está dado por un factor δ multiplicado por el número de residuos con ángulos de torsión en la cadena lateral. Por ejemplo, si se considera $\delta = 50$ y la proteína tiene 300 residuos con ángulos de torsión en la cadena lateral, entonces el número máximo de evaluaciones de la FS es $50 \times 300 = 15000$.

En cuanto a los conjuntos de prueba utilizados, a partir de los resultados de la comparación de los métodos del estado del arte que se muestran en la Tabla 4, fueron elegidos el Dataset-65 y el Dataset-360; ya que en cuanto a precisión son el conjunto más fácil y más difícil de predecir, respectivamente.

El análisis realizado en el presente trabajo contiene dos componentes:

- El análisis de convergencia: donde el interés es el análisis del comportamiento de los algoritmos de búsqueda seleccionados en función al número permitido de evaluaciones de la función de score, lo cual viene dado por el factor constante δ. Los valores seleccionados para δ fueron: 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000 y 100000.
- Comparación de desempeño: empleando las funciones de score de CIS-RR y RASP, el objetivo es obtener al menos la precisión alcanzada por los métodos del estado del arte.

Como todos los algoritmos de búsqueda seleccionados son estocásticos, se realizaron 10 ejecuciones para cada configuración (algoritmo de búsqueda, δ, FS y proteína). La Figura 32 muestra los promedios de scores para cada método de búsqueda y conjunto de prueba, empleando las funciones de score de CIS-RR (parte superior) y RASP (parte inferior). Un punto interesante es que las gráficas siguen la misma tendencia: no existe un método que sea el mejor para todos los casos. Sin embargo, el SAOPUS aparece como el algoritmo de búsqueda más consistente, seguido por el ILS y el SARL. Los valores de score inicialmente altos del SA podrían deberse a un posible valor alto de la temperatura inicial y a un bajo número de iteraciones. No obstante, a partir de $\delta = 5000$ el desempeño de los métodos seleccionados es similar. No se realizaron cambios subsecuentes en los parámetros de cada método de búsqueda ya que puede notarse que el desempeño de cada método se aproxima a los demás a medida que crece el valor de δ. Este comportamiento es de hecho esperado para grandes valores de δ (Aarts y Lenstra, 2003), pero puede verse en la Figura 33 que el tiempo promedio de búsqueda necesario para $\delta = 5000$ no es muy alto (casi dos segundos en promedio para una proteína).

Las precisiones obtenidas por estos algoritmos de búsqueda se muestran en las Ta-

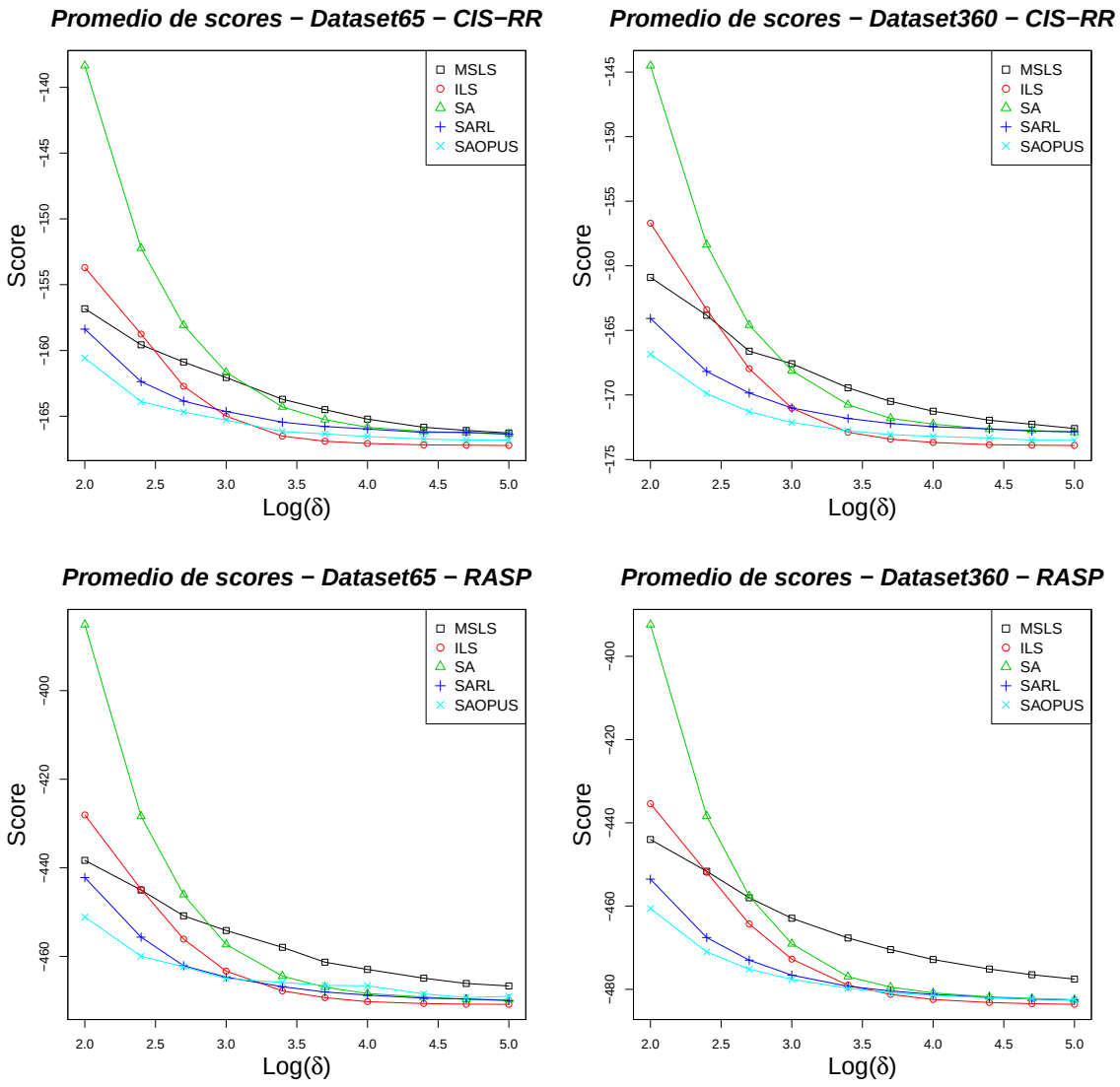

Figura 32. Valores promedio considerando las funciones de score de CIS-RR (parte superior) y RASP (parte inferior) para cada algoritmo de búsqueda y cada conjunto de prueba.

blas 13 y 14, donde también aparecen como referencias las precisiones obtenidas por los métodos del estado del arte. Puede notarse que las precisiones obtenidas con las funciones de score de CIS-RR y RASP, junto con los algoritmos basados en la búsqueda local, compiten con los métodos dados como referencia. Otro punto interesante es que se usó la misma biblioteca de rotámeros y el mismo algoritmo de búsqueda que en OPUS-Rota (Lu *et al.*, 2008a); lo cual indica claramente que su FS es la razón principal de sus mejores resultados de precisión en comparación al resto de los métodos del estado del arte seleccionados y a los algoritmos implementados en este trabajo.

A partir de los resultados del análisis de estas dos funciones de score seleccio-

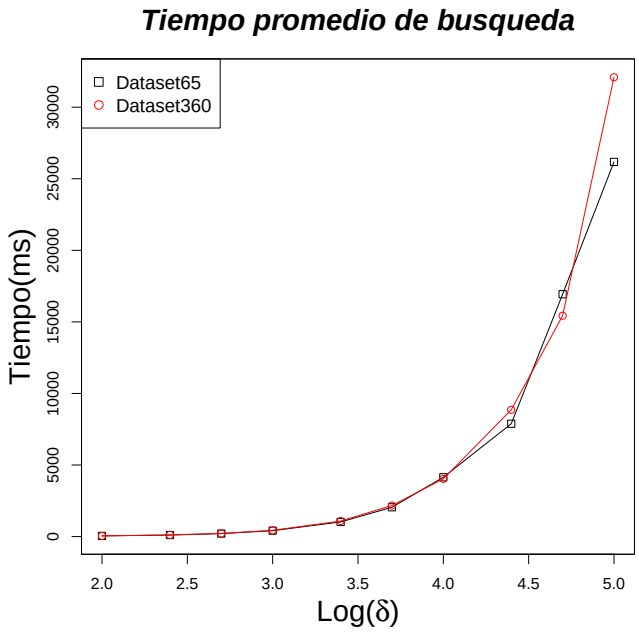

Figura 33. Tiempo promedio para cada conjunto de prueba. Como el número de evaluaciones de cada método de búsqueda local es el mismo y las interacciones se calculan de antemano (ver Subsección 4.1.3), se tienen tiempos de búsqueda idealmente iguales.

nadas, en el siguiente capítulo se plantea y analiza dos posibilidades que expliquen su desempeño. Primero, se explora la posibilidad de una asignación incorrecta de pesos a los términos de las funciones de score a través de un enfoque de optimización biobjetivo. Segundo, dado que las estructuras cristalizadas se emplean generalmente como referencia para la evaluación de los métodos para el PSCPP y la mayoría de estos métodos no consideran los contactos simétricos, se realiza una relajación de las estructuras de referencia mediante dinámica molecular. Con ésto se pretende simular la proteína en un ambiente más realista y evaluar los métodos y/o sus funciones de score de manera más justa.

Tabla 13. Desempeño en términos de la precisión total - Dataset-65

Precisión total con la función de score de CIS-RR

	MSLS		ILS		SA		SARL		SAOPUS	
δ	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$
100	85.75	75.36	85.57	75.04	84.96	72.41	85.85	75.31	**86.39**	**76.11**
250	86.28	76.09	86.16	75.95	85.99	74.85	86.49	76.17	**86.78**	**76.76**
500	86.31	75.97	86.63	76.59	86.53	75.86	86.66	76.45	**86.91**	**76.95**
1000	86.41	76.33	86.83	76.97	86.67	76.36	86.87	76.80	**86.95**	**76.98**
2500	86.63	76.59	**87.08**	**77.23**	86.90	76.72	86.97	76.99	86.99	77.13
5000	86.67	76.77	**87.10**	**77.27**	86.98	76.98	86.98	76.90	87.07	77.22
10000	86.83	76.98	**87.08**	**77.25**	87.04	77.09	87.05	77.03	87.07	77.14
25000	86.97	77.07	**87.12**	**77.30**	87.02	77.03	87.06	77.13	87.08	77.21
50000	86.94	77.07	87.10	77.27	87.05	77.03	87.13	77.16	**87.12**	**77.31**
100000	87.03	**77.29**	**87.11**	**77.29**	87.05	77.11	87.09	77.14	**87.11**	77.22

Precisión total con la función de score de RASP

	MSLS		ILS		SA		SARL		SAOPUS	
δ	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$
100	86.27	75.64	85.83	75.14	85.35	73.34	86.32	75.78	**86.99**	**76.79**
250	86.47	76.02	86.43	76.00	86.56	75.84	86.82	76.72	**87.37**	**77.44**
500	86.71	76.46	86.98	77.00	87.11	76.82	87.19	77.29	**87.40**	**77.59**
1000	86.82	76.70	87.33	77.46	87.27	77.23	87.44	77.61	**87.50**	**77.73**
2500	87.14	77.07	87.55	77.84	87.47	77.75	87.43	77.68	**87.58**	**77.86**
5000	87.22	77.30	87.61	**78.01**	87.52	77.86	87.57	77.85	**87.62**	77.93
10000	87.27	77.38	87.64	**78.03**	87.60	77.92	**87.65**	77.97	87.62	77.94
25000	87.41	77.58	87.68	**78.05**	87.68	**78.05**	**87.70**	78.04	87.58	77.95
50000	87.43	77.64	87.67	78.06	**87.70**	**78.11**	87.65	78.04	87.66	78.03
100000	87.55	77.75	87.66	78.04	87.66	**78.12**	**87.70**	78.08	87.68	78.04

Precisión total - Estado del arte (Tabla 4)

	$\chi_1(\%)$	$\chi_{1+2}(\%)$
SCWRL4	86.45	77.18
OPUS-Rota	**88.75**	**80.46**
CISRR	87.07	78.07
RASP	86.63	75.54
SIDEpro	87.97	79.02

Tabla 14. Desempeño en términos de la precisión total - Dataset-360

Precisión total con la función de score de CIS-RR

	MSLS		ILS		SA		SARL		SAOPUS	
δ	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$
100	84.49	73.49	84.21	73.13	83.94	71.30	84.73	73.75	**85.18**	**74.50**
250	84.77	73.94	84.91	74.12	85.01	73.55	85.34	74.69	**85.62**	**75.11**
500	85.11	74.40	85.41	74.80	85.49	74.65	85.59	75.06	**85.76**	**75.35**
1000	85.25	74.56	85.82	75.38	85.75	75.14	85.79	75.33	**85.90**	**75.57**
2500	85.48	74.94	**86.09**	**75.82**	85.92	75.48	85.91	75.54	86.02	75.69
5000	85.68	75.22	**86.17**	**75.94**	86.00	75.66	85.96	75.62	86.10	75.84
10000	85.80	75.41	**86.19**	**75.97**	86.04	75.70	85.99	75.68	86.09	75.83
25000	85.90	75.56	**86.20**	**75.98**	86.06	75.77	86.00	75.71	86.12	75.90
50000	85.93	75.61	**86.20**	**75.99**	86.06	75.78	86.07	75.77	86.13	75.90
100000	86.02	75.76	**86.20**	**75.99**	86.08	75.78	86.07	75.78	86.13	75.89

Precisión total con la función de score de RASP

	MSLS		ILS		SA		SARL		SAOPUS	
δ	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$	$\chi_1(\%)$	$\chi_{1+2}(\%)$
100	84.64	73.72	84.40	73.38	84.30	71.96	85.07	74.23	**85.51**	**75.05**
250	84.96	74.19	85.04	74.29	85.41	74.28	85.61	75.11	**85.92**	**75.67**
500	85.17	74.46	85.53	75.03	85.80	75.17	85.89	75.55	**86.07**	**75.85**
1000	85.43	74.80	85.95	75.60	86.07	75.69	86.09	75.85	**86.15**	**76.00**
2500	85.63	75.17	86.22	76.04	**86.24**	**76.07**	86.22	76.01	86.22	**76.07**
5000	85.74	75.34	**86.33**	**76.21**	86.30	76.13	86.25	76.15	86.30	76.17
10000	85.90	75.54	**86.39**	**76.32**	86.33	76.22	86.31	76.19	86.33	76.23
25000	86.00	75.72	**86.42**	**76.36**	86.36	76.24	86.34	76.22	86.35	76.25
50000	86.07	75.83	**86.43**	**76.36**	86.37	76.27	86.38	76.27	86.37	76.28
100000	86.17	75.93	**86.44**	**76.39**	86.41	76.33	86.39	76.30	86.37	76.29

Precisión total - Estado del arte (Tabla 4)

	$\chi_1(\%)$	$\chi_{1+2}(\%)$
SCWRL4	86.45	77.05
OPUS-Rota	**87.49**	**78.20**
CISRR	86.11	76.47
RASP	85.86	75.09
SIDEpro	86.77	77.65

Capítulo 5. Análisis de pesos en funciones de score y simulaciones de dinámica molecular a estructuras de referencia

Los resultados del capítulo anterior indican que las funciones de score son las principales responsables por la brecha existente entre la máxima precisión alcanzable y las obtenidas por los métodos del estado del arte. En este capítulo se analizan dos posibles explicaciones para las limitaciones de las funciones de score actuales para el PSCPP: (i) una incorrecta asignación de pesos a los términos de la función de score, o (ii) la conformación restringida de las estructuras cristalográficas empleadas como referencia durante la evaluación de los resultados de predicción. Para responder estos planteamientos: (i) se modela el PSCPP como un problema de optimización biobjetivo, optimizando, al mismo tiempo, los dos términos más importantes de las dos funciones de score evaluadas en el capítulo anterior; y (ii) se realiza un pre-procesamiento de relajación de las estructuras de referencia obtenidas mediante cristalografía de rayos X, a través de simulaciones de dinámica molecular para simular la proteína en el solvente. La mayor parte del contenido de este capítulo se basa en un artículo de revista producto de este trabajo de investigación (Colbes *et al.*, 2018).

5.1. El PSCPP como un problema de optimización multiobjetivo

Las funciones de score se definen generalmente como una suma ponderada de sus términos[1], y la importancia relativa asignada a cada uno de ellos es una decisión crucial durante el proceso de diseño de la función de score (Guerois *et al.*, 2002; Li *et al.*, 2013). Al analizar los resultados del capítulo anterior surge la siguiente interrogante: si la función de score es una suma ponderada de términos, ¿podría darse el caso de que las limitaciones de las funciones de score actuales para alcanzar mejores resultados de precisión se deban a una incorrecta asignación de los pesos? Para responder esta pregunta se propone modelar el PSCPP como un problema de optimización multiobjetivo.

El razonamiento detrás de esta propuesta es el siguiente: los términos empleados

[1] Estos términos usualmente representan las interacciones (físicas y/o basadas en el conocimiento) entre residuos de la proteína, considerando también el ambiente en el que se encuentra la misma.

en las funciones de score de los métodos actuales están usualmente en conflicto entre ellos; lo cual significa que si se considera una FS particular y se la llama F, entonces no es posible minimizar todos sus términos simultáneamente (Miettinen, 2012). Por lo tanto, en lugar de una solución óptima, existe un conjunto de soluciones de compromiso entre los términos en conflicto de F. Una vez que se calculen las mejores soluciones de compromiso (cuya definición se da más adelante) para los términos de F para un caso del PSCPP, se necesita verificar si la estructura más similar a la experimental ($Best$[2]) se encuentra dentro de la curva de las mejores soluciones de compromiso. Si no está, esto significa que independientemente de la combinación lineal de términos que se decida usar, siempre habrá una estructura con menor precisión pero con mejor score que la estructura similar a la experimental (Geoffrion, 1968). Las funciones de score de CIS-RR y RASP se examinan nuevamente para determinar si las estructuras similares a la experimental están presentes en el conjunto de mejores soluciones de compromiso.

Un problema de optimización multiobjetivo puede definirse, sin pérdida de generalidad, de la siguiente manera:

$$\begin{aligned} \text{minimizar} \quad & F(x) = (f_1(x), \ldots, f_m(x)) \\ \text{sujeto a} \quad & x \in X, \end{aligned} \tag{18}$$

donde X es el *espacio de decisión* o *espacio de variables*, R^m es *espacio objetivo* y $F : X \to R^m$ consiste de m funciones objetivo de valores reales (Li y Zhang, 2009). En el contexto de este trabajo, estas funciones objetivo son los términos de las funciones de score para el PSCPP.

Considerando dos vectores $u, v \in R^m$, u *domina* a v si $u_i \leq v_i$ para todo $i = 1, \ldots, m$ y $u \neq v$. Un punto $x^* \in X$ es *óptimo de Pareto* si no existe $x \in X$ tal que $F(x)$ domine a $F(x^*)$. El *conjunto Pareto* (CP) es el conjunto de todos los puntos óptimos de Pareto, y el *frente Pareto* $FP = \{F(x) \in R^m | x \in CP\}$ es el conjunto de todos los vectores objetivo de Pareto (Li y Zhang, 2009), como se muestra en la Figura 34. Por lo tanto, la solución al problema de optimización definido más arriba ya no es una solución única, sino un conjunto de soluciones que están en conflicto entre sí. Si se tiene un algoritmo que pudiese obtener el conjunto Pareto, entonces sería necesario asegurarse de que la

[2] El proceso de obtención de la estructura *Best* se encuentra en la Sección 3.8.

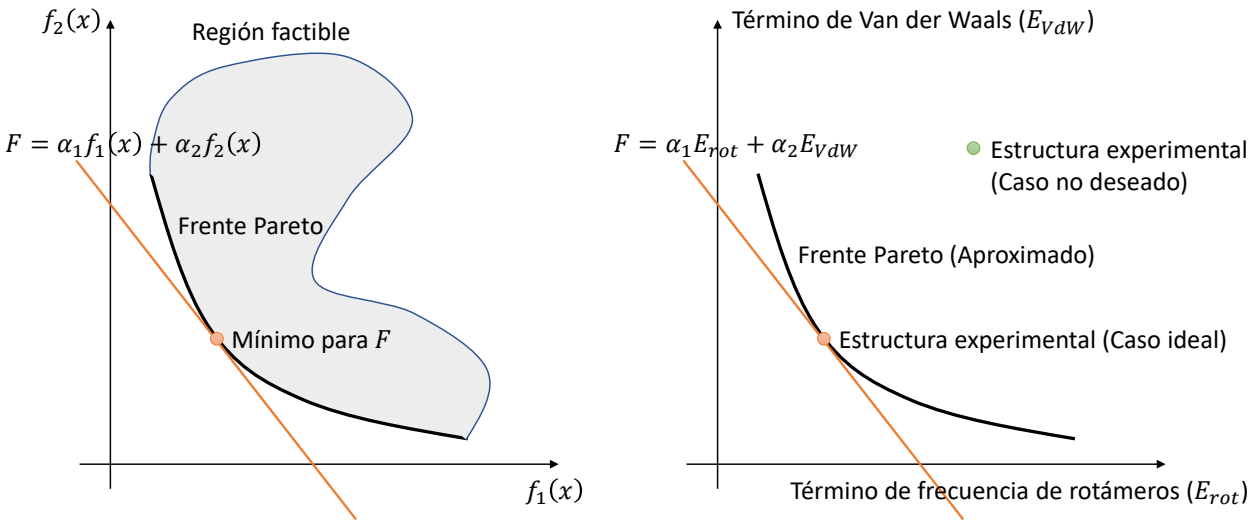

Figura 34. El PSCPP como problema de optimización biobjetivo. *Izq*: Frente Pareto de un problema de optimización multiobjetivo. El enfoque de la suma ponderada (*F*) describe una tangente al frente Pareto en un punto correspondiente al mínimo para *F*. *Der*: Existe una posibilidad de encontrar pesos bajo los cuales la estructura experimental (o una muy *similar*) sea el mínimo para *F* solamente si la estructura experimental está en el frente Pareto.

estructura experimental será parte de este conjunto. Si éste no es el caso, entonces los algoritmos para el PSCPP no podrán devolver como salida la estructura experimental (o una muy parecida a ella), pues siempre encontrarán otras mejores soluciones en términos de la función de score pero con una peor precisión en términos de $\chi_1(\%)$ y $\chi_{1+2}(\%)$.

Una manera de resolver problemas de optimización multiobjetivo consiste en la *escalarización* o la *técnica de la suma ponderada* (Caramia y Dell'Olmo, 2008), la cual coincidentemente es la forma en la que se define la mayoría de las funciones de score para el PSCPP. En este enfoque:

$$\begin{aligned} &\text{minimizar } F = \sum_{i=1}^{m} \alpha_i f_i(x) \\ &\alpha_i > 0, i = 1, \ldots, m \end{aligned} \tag{19}$$

donde α_i es el peso asignado a la función objetivo $f_i(x)$. La Figura 34 ilustra este concepto para dos objetivos, donde *F* representa una línea que es tangente al frente Pareto. Por lo tanto, una *F* ideal para el PSCPP debe ser tangente al frente Pareto en un punto correspondiente a la estructura experimental. Pero si el punto correspondiente a esta estructura no está en el frente Pareto, entonces es imposible que esta estructura

sea un mínimo para *F* bajo cualquier combinación lineal de sus términos (Geoffrion, 1968). Encontrar el conjunto Pareto no es una tarea sencilla. Afortunadamente, se puede evitar este cálculo si se encuentra alguna solución (no necesariamente en el conjunto Pareto), con peor precisión, que domine a la estructura experimental; ya que esto implica que la estructura experimental no puede ser parte del conjunto Pareto.

En la mayoría de los métodos del estado del arte, el término correspondiente a las interacciones de Van der Waals y el término derivado de las probabilidades de los rotámeros en la biblioteca (término de frecuencias de rotámeros) son los que tienen mayores pesos en sus funciones de score. Además, los autores de OPUS-Rota (Lu *et al.*, 2008a) analizaron la contribución de cada término en su FS, concluyendo que los términos de Van der Waals y el de las frecuencias de rotámeros son los componentes más importantes. En este trabajo también se realizó un análisis similar para las funciones de score de CIS-RR (Cao *et al.*, 2011) y RASP (Miao *et al.*, 2011) (los resultados no se muestran), llegando a la misma conclusión. Por lo tanto, se seleccionaron estos términos como las funciones objetivo a minimizar.

Estos objetivos se encuentran usualmente en conflicto entre sí: el término de frecuencias de rotámeros se basa a lo más únicamente en los ángulos de torsión de la cadena principal y en la probabilidad de los rotámeros en la biblioteca, mientras que el término de Van der Waals toma en cuenta las interacciones entre átomos cercanos de la proteína (independientemente de la posición, en la secuencia de la proteína, de los residuos a los que corresponden). La mayoría de los cinco métodos del estado del arte seleccionados en el presente trabajo inician con el rotámero de mayor probabilidad para cada residuo, y mientras esto claramente minimiza el término relacionado a las frecuencias de rotámeros, el score de Van der Waals es generalmente alto. Lo mismo ocurre cuando se minimiza únicamente el término de Van der Waals. La suma ponderada usualmente provee mejores resultados que minimizando solamente un término en particular (Lu *et al.*, 2008a).

El interés principal de esta parte del presente trabajo no está en encontrar el conjunto Pareto para el PSCPP; sino más bien en determinar si la mejor[3] estructura que se puede construir mediante los rotámeros en una biblioteca dada está presente en un conjunto *no dominado* de soluciones aproximadas para el PSCPP. Si no está, enton-

[3]La estructura *Best* definida en la Sección 3.8.

ces tampoco estará en el conjunto Pareto. Se utiliza un enfoque simplista (*naive*) para determinar un conjunto no-dominado de soluciones, y éstas se obtienen mediante los algoritmos basados en la búsqueda local definidos en la Sección 4.3.

Se debe considerar que si se quiere evaluar el score de una estructura experimental en cuanto al término relacionado a las frecuencias de rotámeros, las conformaciones de las cadenas laterales de los residuos de la estructura experimental no se encuentran necesariamente entre los rotámeros de la biblioteca; y por lo tanto no tienen una probabilidad que permita calcular el score. Para resolver esta situación, se asignó una probabilidad *ficticia* a la conformación experimental de cada residuo, determinada por la correspondiente al rotámero más *cercano*[4] en la biblioteca.

Para cada proteína de cada conjunto de prueba (Dataset-65 y Dataset-360) y para cada función de score, se consideró las siguientes estructuras:

- Las estructuras *Experimental*, *Best* y *Best+*; definidas en la Sección 3.8.
- Las soluciones devueltas por SCWRL4, OPUS-Rota, CIS-RR, RASP y SIDEpro.
- Las soluciones de cada uno de los cinco algoritmos basados en la búsqueda local definidos en la Sección 4.3, con los siguientes valores de δ: 100, 250, 500, 1000, 2500 y 5000. Como todos estos métodos son estocásticos, se realizaron 10 ejecuciones para cada valor de δ.

Si las estructuras *Experimental*, *Best* o *Best+*, están en el frente Pareto, entonces existirá una combinación lineal de pesos que permite llegar a estas soluciones, es solo cuestión de encontrarla. Sin embargo, si estas soluciones son dominadas por alguna otra con peor precisión, entonces no existirá combinación lineal de pesos alguna que pueda llevar al algoritmo hacia ellas.

La estructura experimental generalmente tiene diferentes longitudes y ángulos entre enlaces que los valores estándar empleados en este trabajo (Engh y Huber, 1991). Por lo tanto, las relaciones de score entre las estructuras candidatas y la experimental podrían deberse principalmente a estas diferencias, y no a la variación en los ángulos de torsión de la cadena lateral entre dichas estructuras. Por esta razón también

[4] Esto fue definido en la Subsección 3.6.1

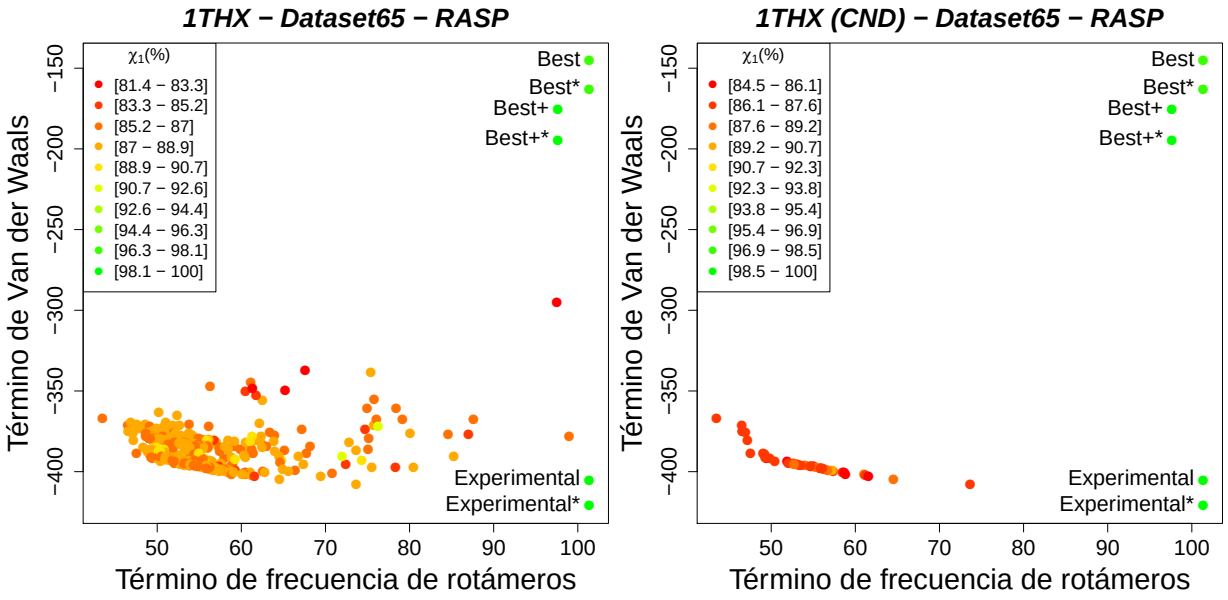

Figura 35. *Izquierda*: scores (en los términos de Van der Waals y de frecuencias de rotámeros del RASP) de diferentes soluciones para la proteína con identificador 1THX, incluyendo la estructura experimental. Los puntos sin etiqueta representan principalmente las soluciones obtenidas por los algoritmos basados en la búsqueda local. El color indica la precisión en $\chi_1(\%)$. *Best* y *Best+* son las soluciones más precisas que pueden obtenerse con los rotámeros de la biblioteca empleada. Las estructuras con * fueron generadas con las longitudes y los ángulos entre enlaces obtenidos de la estructura experimental. *Derecha*: el conjunto no-dominado (CND) de soluciones, y las estructuras *Experimental*, *Best* y *Best+*; para 1THX. Para este caso, y de hecho para casi todas las proteínas de los conjuntos de prueba (independientemente de la FS), *Best* y *Best+* no se encuentran en el CND de soluciones (incluso con los datos experimentales empleados para generar *Best** y *Best+**).

se consideran las estructuras candidatas construidas con longitudes y ángulos entre enlaces obtenidos de la estructura experimental, para evaluar la influencia de los parámetros estándar empleados. De esta forma se analiza la presencia de las siguientes seis estructuras en el conjunto no dominado de soluciones:

- *Experimental*, *Best* y *Best+*: construidas con longitudes y ángulos entre enlaces propuestos por Engh y Huber (1991).
- *Experimental**, *Best** y *Best+**: construidas con longitudes y ángulos entre enlaces de la estructura experimental. Es importante resaltar que en un escenario real no se tendrá acceso a estos valores.

La Figura 35 muestra un ejemplo de los resultados obtenidos. Este es un caso típico, donde la estructura experimental tiene el menor valor en el término de Van der Waals, por lo que está presente en el conjunto no dominado de soluciones. Sin embargo, las

soluciones más precisas que pueden obtenerse con los rotámeros de la biblioteca (*Best* y *Best+*) son dominadas por otras soluciones con peor calidad en χ_1(%). Además, los resultados indican que inclusive si se construyen estas dos soluciones particulares con los parámetros de la estructura experimental (*Best** y *Best+**), éstas siguen dominadas por otras soluciones menos precisas.

Tabla 15. Número de apariciones para cada una de las seis estructuras en el conjunto no dominado de soluciones

	Dataset-65		Dataset-360	
	CIS-RR	RASP	CIS-RR	RASP
Experimental*	56	53	282	279
Experimental	33	31	207	197
Best*	0	0	0	0
Best	0	0	0	0
Best+*	0	0	0	0
Best+	0	0	0	0

La Tabla 15 indica el número de apariciones de estas seis estructuras dentro del conjunto no dominado de soluciones, para cada conjunto de prueba y cada función de score. En el resto de las proteínas de cada conjunto de prueba, estas estructuras son dominadas por otras soluciones; por lo que no pueden estar presentes en el frente Pareto. Puede verse que cuatro de estas estructuras (*Best*, *Best**, *Best+*, and *Best+**), a pesar de ser muy similares a la experimental, no están presentes en el conjunto no dominado para cualquier proteína y FS. Por lo tanto, esto indica que los pesos asignados a cada término de la función de score no serán relevantes para obtener mejores resultados para el PSCPP.

Considerando que este análisis, y las conclusiones que se obtienen mediante él, dependen de la biblioteca de rotámeros seleccionada; se ha agregado la conformación experimental de cada residuo a su conjunto de posibles rotámeros, repitiéndose todo el proceso. Como se añaden las conformaciones de las cadenas laterales de la estructura experimental a la biblioteca en este nuevo escenario, se tiene que *Best* y *Best+* coinciden con la estructura experimental. Al comparar las Figuras 35 y 36, puede observarse que muchas de las soluciones obtenidas por los algoritmos basados en la búsqueda local ahora dominan a la estructura experimental, notándose una disminución significativa del score de Van der Waals. Esto indica que ciertas conformaciones experimentales pueden cambiarse por otras, pertenecientes a la biblioteca, para dis-

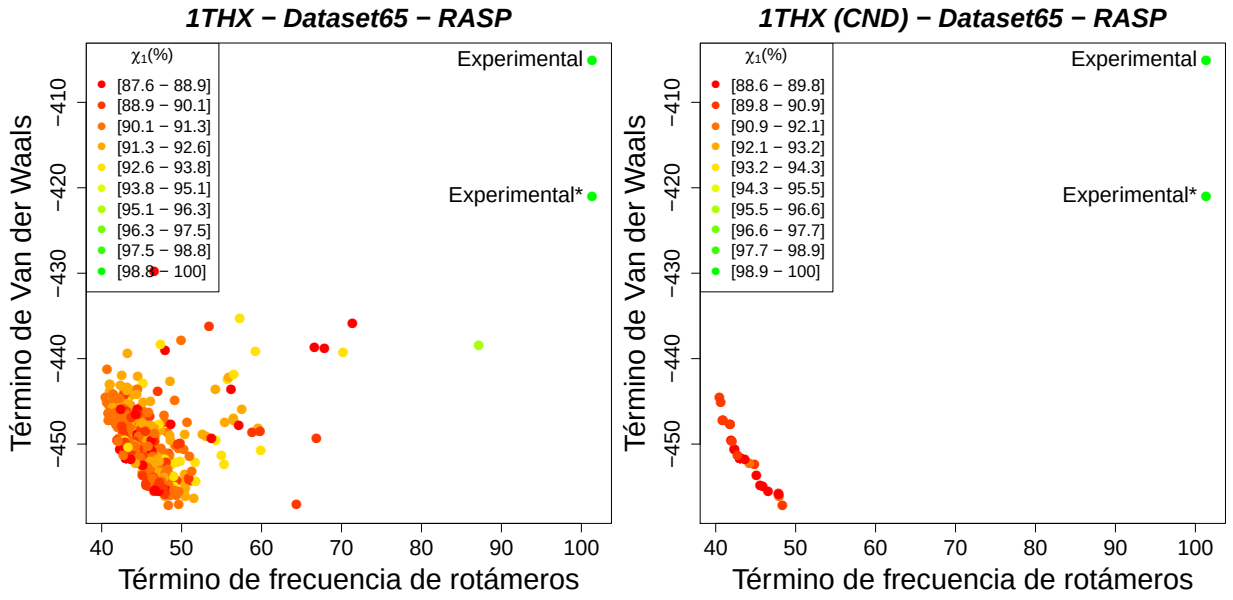

Figura 36. Resultados (explicados en la Figura 35) para el caso de añadir la conformación experimental de un residuo a su conjunto de rotámeros posibles. De esta manera, *Best* y *Best+* se convierten en la estructura experimental. Puede observarse que la estructura experimental es dominada por varias soluciones obtenidas con esta versión *extendida* de la biblioteca de rotámeros. A diferencia de la Figura 35, existen varias soluciones con menores valores para el término de Van der Waals que el del correspondiente a la estructura experimental. Esto indica que algunas de las conformaciones de la cadena lateral en la estructura experimental se pueden cambiar por otras presentes en la biblioteca, para mejorar los valores para este término.

minuir aún más los valores para el término de Van der Waals. De la misma manera que la Tabla 15, la Tabla 16 presenta el número de apariciones de las estructuras *Experimental* y *Experimental** dentro del conjunto no dominado de soluciones para este nuevo escenario. La comparación de los resultados de estas tablas sugiere que, cuanto más se aproximen los rotámeros de la biblioteca a la conformación experimental de un caso para el PSCPP, es más probable que la estructura experimental no se encuentre en el frente Pareto; lo cual da un mayor soporte a la conclusión principal establecida en el párrafo anterior.

5.2. Dinámica molecular para estructuras cristalográficas de referencia

Los conjuntos de prueba para la evaluación de los métodos del estado del arte para el PSCPP usualmente contienen estructuras determinadas por cristalografía de rayos X. De hecho, todas las estructuras de las proteínas en los conjuntos de prueba empleados en este trabajo se obtuvieron mediante esta técnica. Por esta razón, los residuos de proteínas cercanas en el cristal (definidos como *contactos simétricos*) deben con-

Tabla 16. Número de apariciones de la estructura experimental en el conjunto no dominado de soluciones al añadir la conformación experimental de un residuo a su conjunto de rotámeros posibles. *Best* y *Best+* se convierten en la estructura experimental en este caso. Las apariciones en el conjunto no dominado de soluciones es menor que en la Tabla 15, y esto se debe a que la presencia de los *rotámeros experimentales* ayuda a obtener scores menores. Esto puede notarse al comparar las Figuras 35 y 36.

	Dataset-65		Dataset-360	
	CIS-RR	RASP	CIS-RR	RASP
Experimental*	2	3	4	9
Experimental	0	0	0	0

siderarse, ya que tienen una importante influencia en la conformación de las cadenas laterales de la estructura experimental (Rodriguez *et al.*, 1998). Esto es especialmente relevante para el caso de las cadenas laterales de residuos polares en la superficie de la proteína (Jacobson *et al.*, 2002). Los resultados obtenidos en el Capítulo 3 indican que la inclusión de los contactos simétricos aumenta significativamente la precisión de los métodos del estado del arte cuando se emplean las estructuras cristalográficas como referencias; por lo que las limitaciones se pueden atribuir a las funciones de score y los algoritmos de búsqueda (Colbes *et al.*, 2016).

Sin embargo, la mayoría de los métodos del estado del arte para el PSCPP emplean únicamente la secuencia de la proteína y la conformación de su cadena principal, sin considerar los contactos simétricos. Esto se debe a que el objetivo principal de estos métodos es predecir el empacamiento de la cadena lateral de la proteína en solución (Jacobson *et al.*, 2002; Lu *et al.*, 2008a). SCWRL4 es el único de los cinco métodos seleccionados en el presente trabajo que permite la consideración de las interacciones entre residuos de proteínas vecinas en el cristal.

Por lo tanto, la situación actual para el PSCPP es la siguiente: la mayoría de los métodos del estado del arte predice el empacamiento de la cadena lateral de una proteína considerando únicamente su secuencia, la conformación de su cadena principal (Lu *et al.*, 2008a; Cao *et al.*, 2011; Miao *et al.*, 2011; Nagata *et al.*, 2012), y en algunos casos la presencia del solvente (Lu *et al.*, 2008a; Koehl *et al.*, 2011); pero los mismos se evalúan con estructuras de referencia que se obtuvieron en condiciones distintas[5]. De hecho, las estructuras obtenidas mediante cristalografía de rayos X pueden no estar en su conformación nativa[6], por lo que tienen una utilidad limitada en términos

[5] Revisar la Subsección 2.2.3.1

[6] Debido a que estas estructuras resultan de las interacciones de múltiples unidades idénticas.

biológicos (Jacobson *et al.*, 2002). Consecuentemente, podría darse el caso de que el entorno determinado por el cristal en las estructuras empleadas como referencia esté influyendo negativamente en la evaluación del desempeño de los métodos para el PSCPP, y específicamente en la evaluación de sus funciones de score.

Por estas razones, se decidió pre-procesar las estructuras determinadas por cristalografía de rayos X mediante dinámica molecular con el *método de recocido simulado* (MD-SA por sus siglas en inglés). Esto permitiría que los residuos con contactos simétricos adopten una configuración distinta, simulando la conformación de la proteína en presencia del solvente. Así, los métodos para el PSCPP (o específicamente sus funciones de score) podrían evaluarse en una situación bajo la cual las proteínas realizan sus funciones, y también se evitaría la presencia de las fuerzas en el empacamiento determinado por la red cristalina. No se consideró el uso de estructuras obtenidas mediante resonancia magnética nuclear (NMR[7]) para la evaluación de métodos para el PSCPP, debido a que las conformaciones de la cadena lateral están pobremente restringidas en los datos experimentales (Jacobson *et al.*, 2002).

Para la simulación de dinámica molecular de una proteína, solamente se consideró la unidad asimétrica de su estructura cristalográfica rodeada de agua. Se seleccionaron 25 proteínas monoméricas[8] del Dataset-65, realizándose las simulaciones MD-SA mediante NAMD (Phillips *et al.*, 2005). La duración de cada simulación fue de 50 ns y las conformaciones se guardaron cada 20 ps. Mayores detalles de la simulación se dan en el Anexo A.5. Se limitó el estudio a 25 proteínas debido al gran esfuerzo computacional requerido para realizar la simulación MD-SA para una proteína.

La gráfica de RMSD (de átomos pesados de la cadena principal) respecto a la estructura al final del proceso de calentamiento del MD-SA fue considerada para la selección de 250 conformaciones guardadas de la simulación de una proteína en particular. Mediante la gráfica de RMSD se identificó visualmente el punto de estabilidad; y a partir de ahí en adelante, se seleccionaron 250 conformaciones (10 % del total) a intervalos constantes. La Figura 37 muestra la gráfica de RMSD para la proteína 2PTH, donde el punto de inicio se definió a los 10 ns. Por lo tanto, en este caso se toma una conformación cada $(50 - 10)/250 = 160$ ps.

[7] Ver la Subsección 2.2.3.
[8] La lista con los identificadores PDB se encuentra en el Anexo A.2.

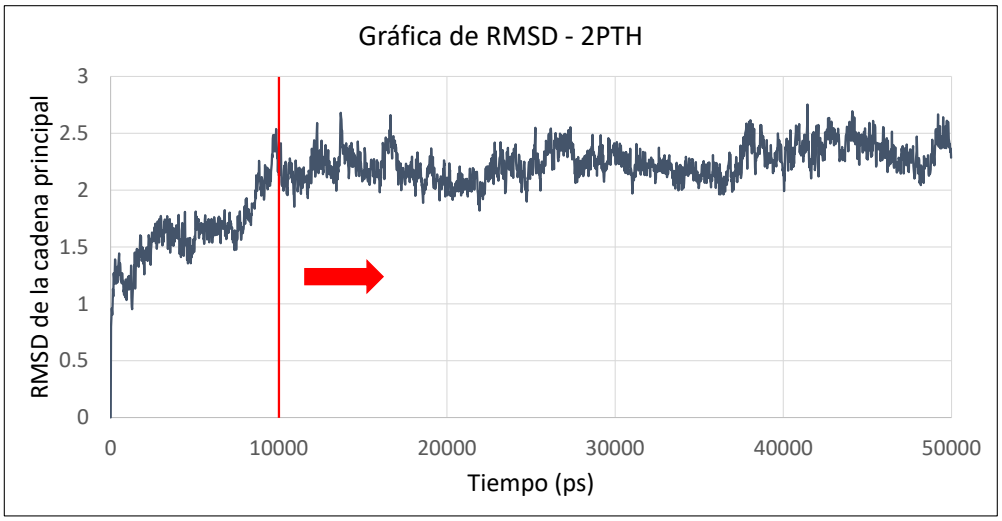

Figura 37. Gráfica de RMSD de la cadena principal en la simulación de dinámica molecular para la proteína con identificador PDB 2PTH. La referencia es la estructura al final del proceso de calentamiento (los detalles se proveen en el Anexo A.5). A partir de la línea vertical roja, se seleccionan 250 conformaciones en intervalos uniformes.

Estas 250 estructuras para cada proteína se emplearon como referencia para evaluar el desempeño de las funciones de score de CIS-RR y RASP, usando el método de búsqueda local definido en el capítulo anterior. Así, cada una de las estructuras seleccionadas a partir de la dinámica molecular se utilizan como entradas para el algoritmo LS, reemplazando a las estructuras cristalográficas que se emplearon anteriormente como referencias.

Con el fin de cuantificar los cambios en las estructuras seleccionadas respecto a la experimental (obtenida por cristalografía de rayos X), se presentan los resultados en las Tablas 17 y 18. La Tabla 17 presenta los datos de accesibilidad al solvente para cada estructura experimental y el promedio para las 250 estructuras seleccionadas de las simulaciones de MD-SA. Se presentan también los cambios de las estructuras experimentales a las de MD entre residuos enterrados y expuestos. Puede observarse que el porcentaje promedio de variación de residuos enterrados a expuestos fue de 7.3 %, mientras que lo opuesto fue de 3.5 %. El cambio de enterrados a expuestos fue mayor que lo opuesto debido a que las proteínas vecinas en el entorno del cristal ya no están presentes en el entorno de la simulación de dinámica molecular.

También se calculó los valores de RMSD, para cada estructura seleccionada, relativo a la experimental correspondiente; y los valores promedio y desviación estándar de los resultados se presentan en la Tabla 18. Los residuos se separan en tres catego-

Tabla 17. Datos sobre accesibilidad al solvente para cada estructura experimental, y el promedio para las 250 estructuras seleccionadas de la dinámica molecular. También se presenta los cambios desde las estructuras experimentales a las estructuras MD entre los residuos enterrados y expuestos.

		Experimental		MD - Prom.		Cambio - Prom.	
PDBID	Num Res	B(%)	E(%)	B(%)	E(%)	B a E(%)	E a B(%)
7RSA	124	38.71	61.29	32.12	67.88	10.32	3.72
1HCL	294	49.32	50.68	43.19	56.81	9.22	3.09
1IFC	131	38.93	61.07	36.12	63.88	5.63	2.82
1MML	251	41.43	58.57	39.52	60.48	6.74	4.82
1CC7	72	26.39	73.61	25.93	74.07	7.08	6.62
2BAA	243	49.38	50.62	42.29	57.71	10.46	3.37
1VJS	469	54.8	45.2	46.25	53.75	11.15	2.59
1AMM	174	41.95	58.05	38.4	61.6	7.69	4.14
1THV	207	43.96	56.04	41.55	58.45	6.85	4.44
1DHN	121	33.88	66.12	33.9	66.1	2.71	2.72
1BD8	156	41.03	58.97	37.12	62.88	7.34	3.44
1IGD	61	19.67	80.33	18.68	81.32	2.44	1.45
1CBN	46	21.74	78.26	16.93	83.07	5.71	0.9
1NAR	289	51.9	48.1	48.13	51.87	7.15	3.38
1WHI	122	40.98	59.02	40.28	59.72	2.1	1.4
1CZ9	139	39.57	60.43	37.24	62.76	5.2	2.87
1EDG	380	54.47	45.53	51.25	48.75	7.96	4.74
1IXH	321	52.65	47.35	47	53	8.48	2.84
1AKO	268	50.75	49.25	47.99	52.01	6.52	3.76
1THX	108	38.89	61.11	34.21	65.79	8.93	4.25
2RN2	155	41.29	58.71	38.23	61.77	6.17	3.11
1NPK	150	42.67	57.33	37.98	62.02	9.77	5.09
1A7S	221	48.42	51.58	42.68	57.32	10.41	4.67
2PTH	193	46.63	53.37	42.45	57.55	8.41	4.23
3LZT	129	41.09	58.91	36.31	63.69	8.94	4.16

B: Enterrado (Buried) - E: expuesto (Exposed).

rías: *Todos*, *Enterrados* y *Expuestos*; mientras que los átomos también se separan en tres categorías: *Todos*, *Cadena principal* y *Cadena lateral*. Como es de esperarse, los valores de RMSD de los residuos enterrados son menores que los correspondientes a los expuestos; y también los valores de RMSD para los átomos de la cadena principal son menores que los correspondientes a los átomos de la cadena lateral.

Se evaluaron las funciones de score de CIS-RR y RASP usando el método de búsqueda local y el conjunto de 25 proteínas. Los tres grupos de estructuras empleadas como referencia fueron:

Tabla 18. RMSD respecto a la estructura experimental para las 250 estructuras seleccionadas de la dinámica molecular para el conjunto de 25 proteínas. Aunque ciertos residuos cambian de enterrados a expuestos (y viceversa), se ignora este hecho debido a los bajos valores presentados en la Tabla 17. Obs: *Todos*: todos los átomos; *BB*: átomos de la cadena principal; *SC*: átomos de la cadena lateral.

RMSD (Promedio y desv. est.)

	Todos los residuos			**Residuos Enterrados**			**Residuos Expuestos**		
PDB ID	Todos	BB	SC	Todos	BB	SC	Todos	BB	SC
7RSA	4.01 ± 0.45	3.46 ± 0.44	4.53 ± 0.47	2.63 ± 0.25	2.67 ± 0.34	2.54 ± 0.17	4.58 ± 0.54	3.83 ± 0.50	5.23 ± 0.59
1HCL	3.14 ± 0.23	2.51 ± 0.25	3.66 ± 0.23	2.11 ± 0.20	1.94 ± 0.21	2.27 ± 0.19	3.81 ± 0.27	2.94 ± 0.29	4.46 ± 0.28
1IFC	1.90 ± 0.08	1.48 ± 0.09	2.23 ± 0.09	1.56 ± 0.08	1.36 ± 0.09	1.73 ± 0.08	2.12 ± 0.10	1.57 ± 0.10	2.54 ± 0.13
1MML	2.62 ± 0.13	2.15 ± 0.16	3.02 ± 0.12	1.89 ± 0.13	1.83 ± 0.16	1.95 ± 0.13	3.06 ± 0.15	2.37 ± 0.18	3.56 ± 0.15
1CC7	2.83 ± 0.36	2.23 ± 0.35	3.32 ± 0.39	2.56 ± 0.37	2.38 ± 0.37	2.70 ± 0.39	2.91 ± 0.37	2.10 ± 0.36	3.49 ± 0.42
2BAA	2.65 ± 0.18	2.25 ± 0.18	3.03 ± 0.18	1.94 ± 0.15	1.93 ± 0.14	1.94 ± 0.18	3.26 ± 0.22	2.58 ± 0.25	3.83 ± 0.22
1VJS	4.90 ± 0.37	4.57 ± 0.37	5.20 ± 0.37	3.39 ± 0.23	3.41 ± 0.22	3.35 ± 0.24	6.25 ± 0.49	5.74 ± 0.52	6.66 ± 0.48
1AMM	2.42 ± 0.15	1.73 ± 0.18	2.89 ± 0.16	1.53 ± 0.17	1.41 ± 0.20	1.63 ± 0.16	2.91 ± 0.16	1.94 ± 0.19	3.49 ± 0.19
1THV	2.91 ± 0.15	2.39 ± 0.19	3.41 ± 0.13	2.11 ± 0.14	2.09 ± 0.16	2.12 ± 0.13	3.54 ± 0.17	2.70 ± 0.22	4.18 ± 0.16
1DHN	3.42 ± 0.16	2.97 ± 0.18	3.79 ± 0.17	2.76 ± 0.13	2.84 ± 0.14	2.61 ± 0.16	3.72 ± 0.20	3.04 ± 0.21	4.19 ± 0.22
1BD8	3.08 ± 0.25	2.47 ± 0.27	3.65 ± 0.26	1.88 ± 0.21	1.84 ± 0.24	1.93 ± 0.17	3.67 ± 0.29	2.90 ± 0.30	4.24 ± 0.30
1IGD	2.74 ± 0.53	2.35 ± 0.58	3.08 ± 0.50	1.26 ± 0.22	1.12 ± 0.20	1.37 ± 0.29	3.01 ± 0.61	2.58 ± 0.68	3.38 ± 0.59
1CBN	4.47 ± 0.43	3.89 ± 0.45	5.02 ± 0.44	2.78 ± 0.33	3.11 ± 0.43	1.40 ± 0.21	4.84 ± 0.48	4.14 ± 0.48	5.41 ± 0.49
1NAR	2.33 ± 0.15	1.82 ± 0.17	2.74 ± 0.14	1.83 ± 0.17	1.57 ± 0.17	2.05 ± 0.19	2.73 ± 0.15	2.02 ± 0.18	3.26 ± 0.15
1WHI	1.56 ± 0.09	1.01 ± 0.09	1.99 ± 0.11	1.01 ± 0.06	1.02 ± 0.08	0.96 ± 0.06	1.81 ± 0.12	0.99 ± 0.12	2.28 ± 0.13
1CZ9	2.53 ± 0.14	2.05 ± 0.14	2.95 ± 0.18	1.60 ± 0.12	1.68 ± 0.17	1.46 ± 0.11	3.01 ± 0.19	2.29 ± 0.17	3.58 ± 0.24
1EDG	3.18 ± 0.10	2.59 ± 0.13	3.68 ± 0.09	2.31 ± 0.11	2.08 ± 0.12	2.52 ± 0.10	3.99 ± 0.12	3.10 ± 0.17	4.70 ± 0.12
1IXH	3.26 ± 0.38	2.95 ± 0.41	3.56 ± 0.35	2.69 ± 0.35	2.64 ± 0.38	2.74 ± 0.32	3.75 ± 0.41	3.23 ± 0.43	4.22 ± 0.41
1AKO	2.75 ± 0.10	1.98 ± 0.11	3.32 ± 0.10	1.98 ± 0.08	1.68 ± 0.11	2.25 ± 0.08	3.28 ± 0.13	2.22 ± 0.12	3.98 ± 0.15
1THX	2.78 ± 0.34	2.14 ± 0.35	3.28 ± 0.36	1.63 ± 0.33	1.54 ± 0.31	1.72 ± 0.36	3.36 ± 0.38	2.50 ± 0.41	3.97 ± 0.41
2RN2	3.08 ± 0.20	2.34 ± 0.20	3.66 ± 0.22	2.06 ± 0.14	1.86 ± 0.13	2.24 ± 0.19	3.64 ± 0.25	2.66 ± 0.27	4.32 ± 0.27
1NPK	3.93 ± 0.22	2.91 ± 0.23	4.79 ± 0.23	2.83 ± 0.21	2.22 ± 0.20	3.39 ± 0.24	4.60 ± 0.26	3.37 ± 0.27	5.53 ± 0.28
1A7S	3.92 ± 0.15	3.36 ± 0.17	4.44 ± 0.15	2.95 ± 0.14	2.80 ± 0.17	3.11 ± 0.12	4.67 ± 0.18	3.94 ± 0.20	5.21 ± 0.19
2PTH	3.36 ± 0.16	2.68 ± 0.16	3.95 ± 0.19	2.48 ± 0.14	2.39 ± 0.16	2.58 ± 0.16	3.96 ± 0.20	2.93 ± 0.21	4.68 ± 0.23
3LZT	4.25 ± 0.48	3.62 ± 0.47	4.80 ± 0.51	3.42 ± 0.53	3.39 ± 0.54	3.42 ± 0.55	4.78 ± 0.48	3.80 ± 0.43	5.48 ± 0.54

- *Dinámica Molecular*: las 250 estructuras elegidas de la simulación MD-SA.
- *Experimental*: las estructuras obtenidas por cristalografía de rayos X y almacenadas en el PDB.
- *WHATIF*: las estructuras resultantes de agregar los contactos simétricos mediante WHATIF (Hekkelman *et al.*, 2010). Para los cálculos de precisión y RMSD, solamente se consideran los residuos de la unidad asimétrica (es decir, se ignoran los residuos agregados por WHATIF).

Para cada una de las conformaciones de referencia de una proteína, se consideraron tres diferentes estructuras como entrada para el algoritmo de búsqueda local:

- *Referencia*: para cada residuo de la proteína, la conformación de la estructura de referencia se agrega a su lista de rotámeros a explorar. La probabilidad de este rotámero agregado es la máxima del conjunto de rotámeros *correctos*[9]
- *Best+* y *Best*: las cuales se definieron en la Sección 3.8. Estas estructuras se basan en la empleada como referencia.

Todas estas estructuras tienen información acerca de la estructura empleada como referencia. El razonamiento principal detrás de este experimento es el siguiente: si se consideran a las conformaciones seleccionadas de la simulación MD-SA como estructuras nativas, entonces las mismas estarían muy cercanas a, al menos, un mínimo local en la función de score evaluada; lo cual implicaría que no se necesitan muchos cambios en los rotámeros para alcanzar estos valores mínimos. Las estructuras *Experimental* y *WHATIF* solamente se emplean para la comparación de resultados.

Por lo tanto, la precisión y las variaciones de score al final del algoritmo LS (ilustrados en la Figura 28) se seleccionaron como los principales indicadores de la calidad de la función de score evaluada. Como se mencionó en el capítulo anterior, el score de una estructura depende del número de residuos de la proteína, por lo que se realiza una normalización dividiendo el score por el número de residuos. Otro indicador importante es el número de proteínas para las que el score al final de la búsqueda local es

[9] Ver su definición en la Subsección 3.6.1.

menor que el correspondiente a la estructura de referencia. Si este número es elevado, entonces claramente la función de score evaluada no garantiza una convergencia a la estructura de referencia a través de un proceso de minimización.

La Tabla 19 muestra los resultados obtenidos en este experimento. Es importante recalcar que las estructuras empleadas como referencia para los cálculos de precisión en los casos *Experimental* y *WHATIF* son las estructuras experimentales obtenidas por cristalografía de rayos X. En el caso de *Dinámica Molecular*, las estructuras empleadas como referencia para los cálculos de precisión son las conformaciones seleccionadas de la simulación MD-SA. Por lo tanto, en este último caso se realizaron 250 ejecuciones del algoritmo de búsqueda local para cada proteína, y se calculó el promedio de los resultados para cada proteína. Los resultados presentados para cada función de score, estructura de referencia y estructura de entrada; son los siguientes:

- La precisión total y el número total de colisiones para las estructuras obtenidas al final de la búsqueda local. La precisión total se calcula por la relación entre el número total de residuos con predicciones correctas[10] y el número total de residuos en el conjunto de 25 proteínas. En el caso de las estructuras de *Dinámica Molecular*, el número total de predicciones correctas para calcular $\chi_1(\%)$ y $\chi_{1+2}(\%)$ es la suma de los promedios de predicciones correctas en las 250 estructuras para cada proteína.
- El promedio de scores al inicio/fin de la búsqueda local. Además, *#Lower* indica el número de proteínas para las cuales la estructura al final del LS tiene un menor score que la estructura de referencia. En el caso de las estructuras de *Dinámica Molecular*, se considera la proporción de estructuras finales (sobre las 250 seleccionadas) con menor score que el correspondiente a la estructura de referencia, para cada una de las 25 proteínas. También se presentan los promedios de scores de las estructuras de referencia.
- El RMSD promedio de la cadena lateral de las estructuras al final de la búsqueda local y el promedio de cambios de rotámeros en el proceso.

Como es de esperarse, los resultados de precisión mejoran cuando se considera

[10] Ver la definición en la Subsección 3.6.1.

Tabla 19. Resultados para las funciones de score de CIS-RR y RASP luego de la búsqueda local, empleando el conjunto de prueba de 25 proteínas y diferentes estructuras de entrada.

Resultados para la función de score de CIS-RR

Precisión total y número de colisiones

	Dinámica Molecular			Experimental			WHATIF		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Referencia	83.7	73.59	0.55	88.88	82.11	2	89.89	83.77	2
Best+	80.33	66.44	6.78	85.76	75.73	10	86.23	76.55	11
Best	80.38	66.52	13.98	85.76	75.73	18	86.21	76.51	18

Promedio de scores y número de estructuras finales con menor score que la referencia.

	MD-SA ($E_{ref} = -0.506$)			Experimental ($E_{ref} = -0.712$)			WHATIF ($E_{ref} = -0.558$)		
	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower
Referencia	-0.506	-0.734	25	-0.712	-0.866	24	-0.558	-0.776	25
Best+	-0.096	-0.581	13.88	-0.262	-0.692	6	-0.065	-0.614	12
Best	-0.096	-0.584	14.191	-0.262	-0.697	6	-0.065	-0.617	12

E_{ref} es el score promedio de las estructuras de referencia, E_{start} es el score de la estructura inicial y E_{end} es el score de la estructura al final del LS.

Promedio del número de cambios de rotámeros y RMSD (promedio y desv. est.)

	Dinámica Molecular		Experimental		WHATIF	
	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios
Referencia	1.575 ± 0.11	101.446	1.264 ± 0.24	86.2	1.231 ± 0.15	85.44
Best+	1.81 ± 0.11	80.224	1.502 ± 0.19	64.96	1.511 ± 0.15	67.8
Best	1.807 ± 0.11	79.823	1.501 ± 0.18	64.68	1.516 ± 0.15	67.56

Resultados para la función de score de RASP

Precisión total y número de colisiones

	Dinámica Molecular			Experimental			WHATIF		
	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.	χ_1(%)	χ_{1+2}(%)	Col.
Referencia	84.16	74.84	2.48	88.97	83.25	4	89.84	84.49	4
Best+	80.47	66.96	6.84	85.71	76.38	11	86.25	77.27	12
Best	80.52	67.05	19.98	85.69	76.45	23	86.23	77.36	28

Promedio de scores y número de estructuras finales con menor score que la referencia.

	MD-SA ($E_{ref} = -0.807$)			Experimental ($E_{ref} = -1.477$)			WHATIF ($E_{ref} = -1.178$)		
	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower	E_{start}	E_{end}	#Lower
Referencia	-0.807	-1.722	25	-1.477	-2.122	25	-1.178	-1.994	25
Best+	0.494	-1.26	18.06	-0.036	-1.632	11	0.429	-1.488	13
Best	0.494	-1.284	18.917	-0.036	-1.652	11	0.429	-1.508	14

E_{ref} es el score promedio de las estructuras de referencia, E_{start} es el score de la estructura inicial y E_{end} es el score de la estructura al final del LS.

Promedio del número de cambios de rotámeros y RMSD (promedio y desv. est.)

	Dinámica Molecular		Experimental		WHATIF	
	RMSD	Cambios	RMSD	Cambios	RMSD	Cambios
Referencia	1.548 ± 0.12	100.162	1.215 ± 0.25	85.16	1.173 ± 0.18	86.08
Best+	1.81 ± 0.12	81.027	1.536 ± 0.22	64.44	1.538 ± 0.16	67.04
Best	1.807 ± 0.12	80.485	1.54 ± 0.22	64.04	1.543 ± 0.16	66.48

Tabla 20. *Arriba*: Desempeño de los cinco métodos del estado del arte en términos de precisión, número de colisiones, RMSD y tiempo de ejecución. Las entradas para estos métodos son las 250 estructuras seleccionadas de la simulación MD-SA para cada una de las 25 proteínas en el conjunto de prueba. *Abajo*: en este caso las entradas con las estructuras cristalográficas, considerando los contactos simétricos. Estos resultados se presentan como referencia, para notar la reducción significativa de la precisión en el caso de las estructuras de MD-SA.

Método	χ_1(%)	χ_{1+2}(%)	Colisiones	RMSD	Tiempo
SCWRL4	76.75	62.69	37	1.963 ± 0.119	1 min 16 s
OPUS-Rota	79.24	65.05	55	1.861 ± 0.123	1 min 31 s
CIS-RR	77.11	63.03	2	1.942 ± 0.113	2 min 11 s
RASP	76.87	61.19	82	1.987 ± 0.117	4 s
SIDEpro	77.2	63.06	6	1.926 ± 0.142	20 s

Método	χ_1(%)	χ_{1+2}(%)	Colisiones	RMSD	Tiempo
SCWRL4	84.47	75.04	32	1.634 ± 0.152	2 min 19 s
OPUS-Rota	86.63	77.4	51	1.481 ± 0.183	2 min 17 s
CIS-RR	85.34	75.66	7	1.545 ± 0.193	3 min 36 s
RASP	84.82	73.14	75	1.674 ± 0.231	4 s
SIDEpro	86.3	76.84	4	1.508 ± 0.219	32 s

los contactos simétricos. Por ejemplo, empezando de las estructuras de referencia, la pérdida de precisión para χ_1(%) disminuye de 11 % a 10 %, aproximadamente. Sin embargo, existe una pérdida significativa de precisión en el caso de las estructuras de *Dinámica Molecular* para todas las estructuras de entrada. Así, al comenzar el proceso de búsqueda local desde las estructuras de referencia, al final del proceso se tiene una pérdida de precisión de aproximadamente 17 % para χ_1(%) (de 100 % a 83 %).

En términos de valores de score, al final de la búsqueda local (con la estructura de referencia como entrada) existe casi siempre una estructura con un menor score que el de la referencia. Esto puede observarse en los valores de *#Lower* para los tres tipos de estructuras de referencia en la Tabla 19. Comenzando desde las estructuras *Best* y *Best+*, casi en la mitad de los casos se tiene un menor score al final de la búsqueda local que el de la referencia (para estructuras *Dinámica Molecular* y *WHATIF*). Como se indicó anteriormente, la razón de este resultado es la conjunción de la forma del paisaje determinado por la función de score y el funcionamiento del algoritmo de la búsqueda local: a medida que la estructura de entrada se aleja de la referencia, existe una mayor probabilidad de estancarse en mínimos locales con menor precisión en relación a la referencia. Además, el RMSD de la cadena lateral aumenta al emplear las estructuras de *Dinámica Molecular* como referencias.

Si se supone que las estructuras de la simulación MD-SA empleadas como referencia se parecen a la nativa de la proteína, los resultados del método de búsqueda local relativos a la pérdida de precisión, el RMSD y los scores; indican que es incluso más difícil alcanzar valores de precisión ideales a través de la minimización del score cambiando el entorno del cristal a uno nuevo que considere únicamente la presencia del solvente. Este resultado implica la existencia de un sesgo hacia estructuras cristalizadas, lo cual se espera dado que las funciones de score y las bibliotecas de rotámeros usualmente se sintonizan mediante estas estructuras. Para reforzar esta observación, se evaluó el desempeño de los cinco métodos del estado del arte seleccionados en el presente trabajo, empleando las estructuras de MD-SA como referencias; y los resultados se presentan en la Tabla 20. Con las estructuras cristalográficas como referencias, el promedio de precisión para estos métodos fue de aproximadamente 85 % para χ_1(%) y 75 % para χ_{1+2}(%); y esta precisión promedio decreció de manera significativa a aproximadamente 77 % para χ_1(%) y 63 % para χ_{1+2}(%) con las estructuras MD como referencias.

Además, bajo estas dos funciones de score evaluadas, casi todas las estructuras de MD-SA tiene scores más altos que el correspondiente a la experimental. La principal preocupación que surge de este hecho es que todas las estructuras seleccionadas tienen menores valores de energía que las estructuras experimentales, bajo la función de energía de CHARMM (Brooks *et al.*, 2009) empleada en la simulación MD-SA. Una posible explicación para esto es el desplazamiento de los átomos de la cadena principal durante la simulación de dinámica molecular, alterando sus distancias con los átomos de la cadena lateral de cada residuo; lo cual a su vez modifica el paisaje inducido por la función de score y los valores asociados con cada punto de este nuevo paisaje. No obstante, también se propone otra posible explicación con base en las observaciones en el presente trabajo: que las funciones de score (y por lo tanto los métodos que las utilizan) para el PSCPP analizadas aquí están sesgadas hacia estructuras obtenidas por cristalografía de rayos X, limitando así el desempeño de los métodos actuales en un entorno más realista como el propuesto en este trabajo.

Luego de observar todos los resultados obtenidos en el presente capítulo, se puede concluir que ni los pesos asignados a las funciones de score ni el entorno del cristal en las estructuras experimentales de referencia son responsables del desempeño insatis-

factorio de las funciones de score actuales para el PSCPP. En el siguiente capítulo se presentan las conclusiones finales y perspectivas de trabajo futuro.

Capítulo 6. Conclusiones

El problema de empacamiento de la cadena lateral en proteínas (PSCPP) es un paso clave tanto en la predicción de estructuras de proteínas como en el diseño de proteínas. Una gran cantidad de métodos para resolver este problema se propusieron en las últimas tres décadas. Sin embargo, desde hace casi una década no se consigue mejorar la precisión de los métodos del estado del arte; y las propuestas recientes se enfocan en otros aspectos tales como el tiempo de ejecución, la facilidad de uso, la exploración de nuevos enfoques, entre otros. Es por ello que en el presente trabajo de tesis se analizaron los tres componentes principales de los métodos para el PSCPP al modelarlo como un problema de optimización combinatoria; con el fin de identificar las limitaciones que impiden obtener mejoras significativas de precisión para este problema. En este capítulo se presenta un resumen con los puntos más importantes tratados durante este trabajo de investigación en la búsqueda de respuestas para la interrogante central en el PSCPP. Posteriormente se establecen las conclusiones principales que resultaron del análisis de los experimentos realizados. Finalmente, se presenta las perspectivas de trabajo futuro que se desprenden de este proyecto de investigación.

6.1. Sumario

Al modelar el PSCPP como un problema de optimización combinatoria se tienen tres componentes principales:

1. La **biblioteca de rotámeros**: que define el espacio de búsqueda.
2. La **función de score**: que indica la calidad de una solución candidata, y es la función objetivo que se desea minimizar.
3. El **algoritmo de búsqueda**: que devuelve soluciones (aproximadas) que consisten en combinaciones de rotámeros que minimicen la función de score.

En este trabajo se realizaron análisis relacionados a cada uno de estos tres componentes, centrándose en la función de score; así como la evaluación de los métodos del estado del arte para el PSCPP. Los puntos principales analizados fueron los siguientes:

1. La evaluación de cinco métodos del estado del arte en términos de precisión, número de colisiones, tiempo de ejecución y RMSD; empleando cuatro conjuntos de prueba que tienen en total 2883 estructuras distintas. Un punto importante es que este trabajo es el primero que realiza una comparación considerando los *contactos simétricos*, los cuales aparecen debido a la determinación experimental de las estructuras de referencia mediante la cristalografía de rayos X.

2. En cuanto a la biblioteca de rotámeros, se calculó la máxima precisión alcanzable con una biblioteca dependiente de la cadena principal, que aparece en cuatro de los cinco métodos del estado del arte seleccionados para la comparación.

3. La evaluación de las funciones de score fue el interés principal en este trabajo. Primero se propuso el método de la búsqueda local, el cual utiliza un algoritmo de búsqueda local con la estructura experimental como entrada. La idea central es que esta estructura debe ser al menos un mínimo local en la función de score evaluada, por lo que la calidad de dicha función está relacionada a la variación de dicha estructura inicial durante la aplicación del algoritmo de búsqueda local. Otros indicadores importantes de este método son el número de cambios de rotámeros en el proceso y la variación total de score. Se implementaron dos funciones de score de los métodos del estado del arte seleccionados (CIS-RR y RASP), para evaluarlos con el método propuesto. En una segunda fase, se exploraron dos posibles causas de las limitaciones de las funciones de score actuales. La primera causa podría ser una incorrecta asignación de los pesos para cada uno de los términos de las funciones de score, por lo cual se estudió el PSCPP desde un enfoque de optimización biobjetivo. La segunda causa podría ser que las condiciones necesarias para la cristalografía de rayos X afecten la conformación nativa de la proteína, y que al utilizarlas como referencias se tendrían resultados engañosos. Por lo tanto, se realizó un relajamiento de las estructuras cristalográficas mediante simulaciones de dinámica molecular; empleando las estructuras resultantes como referencias para el método de evaluación de funciones de score propuesto en este trabajo.

4. En cuanto al algoritmo de búsqueda, se realizaron experimentos para determinar si los algoritmos basados en la búsqueda local se pueden considerar como métodos estándar para el PSCPP. Se analizó la convergencia de resultados y el tiempo

de ejecución necesario para ella.

6.2. Conclusiones

Los resultados de la comparación entre los cinco métodos del estado del arte para el PSCPP seleccionados reveló que los mismos tienen niveles de desempeño similares, obteniendo en promedio una precisión de aproximadamente 87 % para χ_1(%) y 77 % para χ_{1+2}(%). Por otro lado, la máxima precisión alcanzable utilizando la biblioteca de rotámeros estándar de Dunbrack y Cohen (1997) es de aproximadamente 98 % para χ_1(%) y 96 % para χ_{1+2}(%). Al separar la máxima precisión alcanzable por tipo de residuo (enterrado o expuesto), se observó que para los residuos enterrados ya se tienen resultados cercanos al máximo alcanzable en relación a χ_1(%), por lo que las limitaciones de los métodos actuales se encuentran principalmente en los residuos expuestos. Además, para ambos tipos de residuos, χ_{1+2}(%) se encuentra a una mayor distancia de los máximos valores alcanzables. Todo esto indica que aún existe un considerable espacio de mejora para el PSCPP.

La brecha entre la máxima precisión alcanzable y la precisión promedio de los métodos actuales, el desempeño similar de dichos métodos y la casi nula mejora de métodos más recientes; indican que existen limitaciones que impiden alcanzar estas máximas precisiones calculadas. Como una biblioteca de rotámeros estándar ya alcanza valores casi ideales, se puede concluir que las bibliotecas de rotámeros no son responsables de la situación actual del PSCPP.

Para evaluar la calidad de una función de score se debe aislar este componente de los otros dos, y esto se logra mediante el método de búsqueda local propuesto aquí. Al compararlo con el método de búsqueda de conformaciones de un solo residuo, se ha determinado que el método de búsqueda local proporciona información más precisa y útil sobre el desempeño de funciones de score. Los resultados de evaluación de las funciones de score de CIS-RR y RASP, dos métodos del estado del arte para el PSCPP, mediante este método indican que en aproximadamente 99.9 % de las proteínas en los conjuntos de prueba, la estructura experimental de referencia no fue la de mínimo score. Además, la pérdida de precisión para χ_1(%) fue mayor a 8 % al final de la búsqueda local; lo cual sugiere que estas funciones de score no garantizan la

convergencia a la estructura experimental (o a la nativa). Como las funciones de score implementadas son similares a las de los demás métodos del estado del arte, y dada la similitud de sus desempeños en general, se conjetura que se tendrían resultados similares al evaluar sus funciones de score mediante el método de búsqueda local propuesto en este trabajo.

Al analizar el tercer componente de los métodos para el PSCPP, el algoritmo de búsqueda, los resultados obtenidos indican que los algoritmos basados en la búsqueda local alcanzan resultados competitivos en relación a los métodos del estado del arte para el PSCPP. Además, el desempeño en términos de minimización de score de los algoritmos seleccionados converge a medida que el número máximo de evaluaciones de la función objetivo (función de score) aumenta. Sin embargo, convenientemente, el tiempo necesario para esta convergencia no es muy alto. Por lo tanto, estos algoritmos se pueden usar con cualquier propuesta futura de función de score y la biblioteca de Dunbrack y Cohen (1997) para tener nuevas propuestas que aborden el PSCPP como problema combinatorio. Los algoritmos más destacados de acuerdo a los experimentos realizados son el recocido simulado sin rechazo y la búsqueda local iterativa.

Todo indica entonces que las limitaciones de los métodos actuales para el PSCPP se deben a las funciones de score, lo cual también ocurre en la predicción de estructura de proteínas y en el diseño de proteínas. Para proporcionar una explicación sobre los resultados negativos obtenidos en la evaluación de las funciones de score, el análisis de las dos posibilidades planteadas arrojó lo siguiente:

1. Los resultados al modelar el PSCPP como un problema biobjetivo, con el empleo de los dos términos más importantes de las funciones de score (el de Van der Waals y el relacionado a la probabilidad de los rotámeros), indican que las estructuras más similares a la experimental que pueden construirse con los rotámeros de la biblioteca son dominadas por otras soluciones; e incluso en muchos casos se da lo mismo con la estructura experimental. Por lo tanto, es imposible tener a cualquiera de estas dos estructuras como mínimo para cualquier combinación de los términos de las dos funciones de score evaluadas; y se conjetura que lo mismo se cumple para las funciones de score de los demás métodos del estado del arte. Esto descarta la posibilidad de que una incorrecta asignación de pesos en los términos de las funciones de score sea la responsable de sus limitaciones

para el PSCPP.

2. A pesar de pre-procesar las estructuras determinadas por cristalografía de rayos X empleando agua como solvente y simulando el sistema con la función de energía de CHARMM, los resultados indican que la calidad de las funciones de score evaluadas disminuye aún más. Esto sugiere fuertemente que las funciones de score analizadas presentan un sesgo hacia estructuras cristalizadas. Las simulaciones de dinámica molecular también muestran que aunque las conformaciones seleccionadas tienen una menor energía total que la estructura experimental (bajo la función de energía de CHARMM), las funciones de score de CIS-RR y RASP retornan valores más altos para las interacciones de la cadena lateral en las conformaciones seleccionadas que en las estructuras experimentales. Para explicar esta contradicción, se podría argumentar que el desplazamiento de la cadena principal durante la simulación de dinámica molecular afecta el score de la cadena lateral; sin embargo, se conjetura que este resultado contradictorio se debe también a las limitaciones de las funciones de score actuales para el PSCPP. Por lo tanto, se puede concluir que la utilización de estructuras cristalográficas como referencias para medir la calidad de predicción de los métodos evaluados tampoco es responsable de las limitaciones para el PSCPP.

En general, estos resultados sirven para reforzar aún más la idea de que todos los esfuerzos para el PSCPP deben enfocarse en el desarrollo de nuevas funciones de score, que logren guiar correctamente a los algoritmos de búsqueda para obtener soluciones con precisiones cercanas a las máximas alcanzables para el PSCPP.

6.3. Trabajo Futuro

Tal como se señaló en las conclusiones, se cuenta actualmente con una biblioteca de rotámeros que permite alcanzar precisiones cercanas a las ideales. Por otro lado, los algoritmos basados en la búsqueda local son suficientemente buenos para obtener soluciones competitivas para el PSCPP. Por lo tanto, el siguiente paso consiste en diseñar nuevas funciones de score que superen las limitaciones señaladas en este trabajo para las empleadas actualmente.

El método de búsqueda de conformaciones de un solo residuo (SRCS) fue considerado como el criterio principal en el diseño de funciones de score para el PSCPP (Liang y Grishin, 2002), buscando minimizar el RMSD promedio entre el rotámero de menor score para cada residuo y la conformación experimental para cada uno de ellos. En el caso de SIDEpro (Nagata *et al.*, 2012), unas redes neuronales fueron entrenadas para obtener valores de interacción entre pares de residuos de acuerdo a sus tipos y conformaciones, empleando aproximaciones de valores de energía de referencia y el método de Monte Carlo basado en cadenas de Markov. Por lo tanto, los parámetros de una función de score candidata pueden optimizarse mediante un determinado criterio de calidad con un conjunto de entrenamiento.

Como criterio de calidad podría emplearse alguno de los devueltos por el método de la búsqueda local propuesto en el presente trabajo. Por ejemplo, podría buscarse la minimización de la variación de precisión al final de la búsqueda local. En cuanto a la representación de una función de score candidata, la misma podría tener una forma predeterminada o bien considerar un conjunto de operaciones posibles y aplicar programación genética. Más detalles se proporcionan en el Anexo A.6.

Otro punto importante es que las bibliotecas de rotámeros se construyen mediante el análisis de estructuras experimentales almacenadas en el PDB, donde justamente se tiene que casi el 90 % de las estructuras fueron obtenidas mediante cristalografía de rayos X. Por lo tanto, así como las funciones de score están sesgadas hacia estructuras cristalográficas, las opciones de conformaciones dadas por las bibliotecas también podrían presentar un sesgo hacia estructuras de este tipo. Para el diseño tanto de funciones de score como de bibliotecas de rotámeros, se podría emplear entonces como referencia un conjunto de conformaciones obtenidas mediante dinámica molecular, o bien considerar estructuras obtenidas mediante resonancia magnética nuclear (NMR); para eliminar así el sesgo hacia estructuras cristalográficas.

De manera a expandir el abanico de funciones de score disponibles, otra propuesta es implementar más funciones de score de los métodos del estado del arte para el PSCPP; para también de esta manera verificar si efectivamente ciertas conclusiones, establecidas a partir del análisis de las funciones de score de CIS-RR y RASP, son aplicables para las demás. Por otro lado, también podrían realizarse simulaciones de dinámica molecular restringiendo o “congelando” los átomos de la cadena principal

de cada proteína del conjunto de prueba, lo cual permitiría libertad de movimiento únicamente a los átomos de la cadena lateral. Así, el método de búsqueda local podría aplicarse a las estructuras resultantes, bajo las funciones de score evaluadas en este trabajo, para determinar si el desplazamiento de la cadena principal durante las simulaciones de dinámica molecular es o no responsable de los resultados obtenidos en el Capítulo 5.

Literatura citada

Aarts, E. y Lenstra, J. (2003). *Local search in combinatorial optimization*. Princeton University Press.

Akutsu, T. (1997). NP-hardness results for protein side-chain packing. *Genome Informatics*, **8**: 180–186.

Alford, R., Leaver-Fay, A., Jeliazkov, J., O'Meara, M., DiMaio, F., Park, H., Shapovalov, M., Renfrew, P., Mulligan, V., Kappel, K., *et al.* (2017). The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*.

Allen, M. *et al.* (2004). Introduction to molecular dynamics simulation. *Computational soft matter: from synthetic polymers to proteins*, **23**: 1–28.

Anfinsen, C. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096): 223–230.

Bai, X., McMullan, G., y Scheres, S. (2015). How cryo-em is revolutionizing structural biology. *Trends in biochemical sciences*, **40**(1): 49–57.

Baker, D. (2010). An exciting but challenging road ahead for computational enzyme design. *Protein science*, **19**(10): 1817–1819.

Baker, D. y Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**(5540): 93–96.

Bale, J., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T., Gonen, T., King, N., *et al.* (2016). Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science*, **353**(6297): 389–394.

Berg, J., Tymoczko, J., y Stryer, L. (2012). *Biochemistry*. W. H. Freeman, seventh edición. 1120 pp.

Berman, H., Westbrook, J., y Z, Z. F. (2000). The protein data bank. *Nucleic Acids Research*, **28**(1): 235–242.

Berman, H. M., Westbrook, J. D., Gabanyi, M. J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., *et al.* (2008). The protein structure initiative structural genomics knowledgebase. *Nucleic acids research*, **37**(suppl_1): D365–D368.

Bhat, T., Sasisekharan, V., y Vijayan, M. (1978). An analysis of side-chain conformation in proteins. *International Journal of Peptide and Protein Research*, **13**(2): 170–184.

Blundell, T., Sibanda, B., Sternberg, M., y Thornton, J. (1987). Knowledge-based prediction of protein structures. *Nature*, **326**(26): 347–352.

Boas, F. y Harbury, P. (2007). Potential energy functions for protein design. *Current Opinion in Structural Biology*, **17**(2): 199–204.

Boyd, S. y Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brizuela, C., Corona, R., Lezcano, C., Rodríguez, D., y Colbes, J. (2015). An experimental analysis of the performance of sidechain packing algorithms. En: *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*. pp. 929–933.

Brooks, B., Brooks, C., MacKerell, A., Nilsson, L., Petrella, R., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R., Post, C., Pu, J., Schaefer, M., Tidor, B., Venable, R., Woodcock, H., Wu, X., Yang, W., York, D., y Karplus, M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, **30**(10): 1545–1614.

C. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *Journal of molecular biology*, **105**(1): 1–12.

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

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.

Caramia, M. y Dell'Olmo, P. (2008). *Multi-objective management in freight logistics: Increasing capacity, service level and safety with optimization algorithms*. Springer Science & Business Media.

Carugo, O. y Argos, P. (1997). Correlation between side chain mobility and conformation in protein structures. *Protein Engineering*, **10**(7): 777–787.

CBN (1970). Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. *European Journal of Biochemistry*, **17**(2): 193–201.

Chan, W., Zhou, A., y Read, R. (2014). Towards engineering hormone-binding globulins as drug delivery agents. *PloS one*, **9**(11): e113402.

Chandrasekaran, R. y Ramachandran, G. (1970). Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. *International Journal of Protein Research*, **2**: 223–233.

Chou, K. (2004). Structural bioinformatics and its impact to biomedical science. *Current medicinal chemistry*, **11**(16): 2105–2134.

Claverie, J. (2000). From bioinformatics to computational biology. *Genome research*, **10**(9): 1277–1279.

Clote, P. y Backofen, R. (2000). *Computational molecular biology: an introduction*. Wiley series in mathematical and computational biology. John Wiley, primera edición. Chichester, New York. 304 pp.

Colbes, J., Corona, R., Lezcano, C., Rodriguez, D., y Brizuela, C. (2016). Protein side-chain packing problem: is there still room for improvement? *Briefings in bioinformatics*, **18**(6): 1033–1043.

Colbes, J., Aguila, S., y Brizuela, C. (2018). Scoring of Side-Chain Packings: An Analysis of Weight Factors and Molecular Dynamics Structures. *Journal of Chemical Information and Modeling*, **58**(2): 443–452.

Coluzza, I. (2017). Computational protein design: a review. *Journal of Physics: Condensed Matter*, **29**(14): 143001.

Comte, P., Vassiliev, S., Houghten, S., y Bruce, D. (2011). Genetic algorithm with alternating selection pressure for protein side-chain packing and pk(a) prediction. *Biosystems*, **105**(3): 263–270.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., y Kollman, P. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, **117**(19): 5179–5197.

Corona, I. (2010). *Analisis Comparativo de dos Heur?sticas para el Problema de Empaquetamiento de la Cadena Lateral en Proteinas*. Tesis de maestría, CICESE.

Craik, D., Fairlie, D., Liras, S., y Price, D. (2013). The future of peptide-based drugs. *Chemical biology & drug design*, **81**(1): 136–147.

Dahiyat, B. y Mayo, S. (1996). Protein design automation. *Protein Science*, **5**(5): 895–903.

Das, R. (2011). Four small puzzles that rosetta doesn't solve. *PLOS One*, **6**(5): e20044.

Davis, L. (1991). *Handbook of genetic algorithms*, Vol. 115. Van Nostrand Reinhold.

Desjarlais, J. y Handel, T. (1995). De novo design of the hydrophobic cores of proteins. *Protein Science*, **4**(10): 2006–2018.

Desmet, J., Maeyer, M., Hazes, B., y Lasters, I. (1992). The dead-end elimination theorem and its use in side-chain positioning. *Nature*, **356**: 539–542.

Dill, K. y MacCallum, J. (2012). The protein-folding problem, 50 years on. *Science*, **338**(6110): 1042–1046.

Dirac, P. (1929). Quantum mechanics of many-electron systems. En: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, Vol. 123, pp. 714–733.

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

Dunbrack, R. y Cohen, F. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, **6**(8): 1661–1681.

Dunbrack, R. y Karplus, M. (1993). Backbone-dependent Rotamer Library for Proteins Application to Side-chain Prediction. *Journal of Molecular Biology*, **230**(2): 543–574.

Dunbrack, R. y Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Structural & Molecular Biology*, **1**(5): 334–340.

Eiben, A. E., Smith, J. E., *et al.* (2003). *Introduction to evolutionary computing*, Vol. 53. Springer.

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

Eyal, E., Najmanovich, R., Mcconkey, B., Edelman, M., y Sobolev, V. (2004). Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *Journal of computational chemistry*, **25**(5): 712–724.

Fraczkiewicz, R. y Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*, **19**(3): 319–333.

Francis-Lyon, P. y Koehl, P. (2014). Protein side-chain modeling with a protein-dependent optimized rotamer library. *Proteins: Structure, Function, and Bioinformatics*, **82**(9): 2000–2017.

Fung, H., Welsh, W., y Floudas, C. (2008). Computational de novo peptide and protein design: Rigid templates versus flexible templates. *Industrial & Engineering Chemistry Research*, **47**(4): 993–1001.

Gaillard, T., Panel, N., y Simonson, T. (2016). Protein sidechain conformation predictions with an MMGBSA energy function. *Proteins: Structure, Function, and Bioinformatics*, **84**(6): 803–819.

Gaines, J., Virrueta, A., Buch, D., Fleishman, S., O'Hern, C., y Regan, L. (2017). Collective repacking reveals that the structures of protein cores are uniquely specified by steric repulsive interactions. *Protein engineering, design & selection: PEDS*, p. 1.

Gainza, P., Roberts, K., Georgiev, I., Lilien, R., Keedy, D., Chen, C., Reza, F., Andersona, A., Richardson, D., Richardson, J., y Donald, B. (2013). OSPREY: Protein design with ensembles, flexibility, and provable algorithms. *Methods in Enzymology*, **523**: 87–107.

Garey, M. y Johnson, D. (1979). *Computers and intractability. A guide to the theory of NP-completeness.*. WH Freeman and Co.

Geoffrion, A. (1968). Proper efficiency and the theory of vector maximization. *Journal of mathematical analysis and applications*, **22**(3): 618–630.

Geyer, C. (1992). Practical markov chain monte carlo. *Statistical Science*, pp. 473–483.

Goodwin, S., McPherson, J., y McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6): 333–351.

Gordon, D. y Mayo, S. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure*, **7**(9): 1089–1098.

Gordon, D., Marshall, S., y Mayo, S. (1999). Energy functions for protein design. *Current Opinion in Structural Biology*, **9**(4): 509–513.

Greene, J. y Supowit, K. (1986). Simulated annealing without rejected moves. *IEEE Transactions on Computer-Aided Design*, **5**(1): 221–228.

Griss, R., Schena, A., Reymond, L., Patiny, L., Werner, D., Tinberg, C., Baker, D., y Johnsson, K. (2014). Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring. *Nature chemical biology*, **10**(7): 598–603.

Guerois, R., Nielsen, J., y Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, **320**(2): 369–387.

Handl, J., Knowles, J., y Lovell, S. (2009). Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, **25**(10): 1271–1279.

Hekkelman, M., te Beek, T., Pettifer, S., Thorne, D., Attwood, T., y Vriend, G. (2010). Wiws: a protein structure bioinformatics web service collection. *Nucleic acids research*, **38**(suppl_2): W719–W723.

Holm, L. y Sander, C. (1992). Fast and simple monte carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins: Structure, Function, and Bioinformatics*, **14**(2): 213–223.

Huang, P., Boyken, S., y Baker, D. (2016). The coming of age of de novo protein design. *Nature*, **537**(7620): 320–327.

Huerta, M., Downing, G., Haseltine, F., Seto, B., y Liu, Y. (2000). NIH working definition of bioinformatics and computational biology. *Bioinformatics Definition Committee, National Institute of Health, Washington DC*.

Humphrey, W., Dalke, A., y Schulten, K. (1996). Vmd: visual molecular dynamics. *Journal of molecular graphics*, **14**(1): 33–38.

Jacobson, M., Friesner, R., Xiang, Z., y Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. *Journal of molecular biology*, **320**(3): 597–608.

James, M. y Sielecki, A. (1983). Structure and refinement of penicillopepsin at 1.8A resolution. *Journal of Molecular Biology*, **163**(2): 299–361.

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

Jones, T. y Thirup, S. (1986). Using known substructures in protein model building and crystallography. *The EMBO Journal*, **5**(4): 819–822.

Khoury, G. A., Smadbeck, J., Kieslich, C. A., y Floudas, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends in biotechnology*, **32**(2): 99–109.

Kirkpatrick, S., Gelatt, C., y Vecchi, M. (1983). Optimization by simulated annealing. *Science*, **220**(4598): 671–680.

Koehl, P. y Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *Journal of Molecular Biology*, **239**(2): 249–275.

Koehl, P., Orland, H., y Delarue, M. (2011). Adapting Poisson-Boltzmann to the self-consistent mean field theory: Application to protein side-chain modeling. *The Journal of Chemical Physics*, **135**(5): 055104–11.

Koza, J. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, **4**(2): 87–112.

Krieger, E., Nabuurs, S., y Vriend, G. (2005). *Homology Modeling*, pp. 509–523. John Wiley & Sons, Inc.

Krivov, G., Shapovalov, M., y Dunbrack, R. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**(4): 778–795.

Lazaridis, T. y Karplus, M. (2000). Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, **10**(2): 139–145.

Lee, C. y Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *Journal of Molecular Biology*, **217**(2): 373–388.

Lennard-Jones, J. E. (1924). On the determination of molecular fields. En: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, Vol. 106, pp. 463–477.

Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, **65**: 44–45.

Levinthal, C. (1969). How to fold graciously. *Mossbauer spectroscopy in biological systems*, **67**: 22–24.

Lezcano, C. (2012). *Problema de empaquetamiento de la cadena lateral de proteinas: Analisis de la calidad de soluciones de una biblioteca de rotameros simple*. Tesis de maestría, Universidad Nacional de Asuncion.

Li, H. y Zhang, Q. (2009). Multiobjective optimization problems with complicated pareto sets, moea/d and nsga-ii. *IEEE Transactions on Evolutionary Computation*, **13**(2): 284–302.

Li, Z., Yang, Y., Zhan, J., Dai, L., y Zhou, Y. (2013). Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. *Annual Review of Biophysics*, **42**: 315–335.

Liang, S. y Grishin, N. (2002). Side-chain modeling with an optimized scoring function. *Protein Science*, **11**(2): 322–331.

Liang, S., Zheng, D., Zhang, C., y Standley, D. (2011a). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*, **27**(20): 2913–2914.

Liang, S., Zhou, Y., Grishin, N., y Standley, D. (2011b). Protein Side Chain Modeling with Orientation-Dependent Atomic Force Fields Derived by Series Expansions. *Journal of Computational Chemistry*, **32**(8): 1680–1686.

Lin, M., Fawzi, N., y Head-Gordon, T. (2007). Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure*, **15**(6): 727–740.

Liu, H. y Chen, Q. (2016). Computational protein design for given backbone: recent progresses in general method-related aspects. *Current Opinion in Structural Biology*, **39**: 89–95.

Liu, Y. y Kuhlman, B. (2006). RosettaDesign server for protein design. *Nucleic Acids Research*, **34**(2): W235–W238.

Loncharich, R., Brooks, B., y Pastor, R. (1992). Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, **32**(5): 523–535.

Lourenço, H., Martin, O., y Stützle, T. (2010). Iterated local search: Framework and applications. En: *Handbook of metaheuristics*. pp. 363–397.

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

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

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., y Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**(6): 1087–1092.

Miao, Z., Cao, Y., y Jiang, T. (2011). RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, **27**(22): 3117–3122.

Miettinen, K. (2012). *Nonlinear multiobjective optimization*, Vol. 12. Springer Science & Business Media.

Molga, M. y Smutnicki, C. (2005). Test functions for optimization needs. Recuperado (31/05/2018) de: http://new.zsd.iiar.pwr.wroc.pl/files/docs/functions.pdf.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., y Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, **12**(4): 345–364.

Moult, J. (2005). A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, **15**(3): 285–289.

Moult, J., Pedersen, J., Judson, R., y Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, **23**(3).

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., y Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round xi. *Proteins: Structure, Function, and Bioinformatics*, **84**(S1): 4–14.

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

Nagata, K., Randall, A., y Baldi, P. (2012). SIDEpro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, **80**(1): 142–153.

Nair, A. (2007). Computational biology & bioinformatics: a gentle overview. *Communications of the Computer Society of India*, **2**.

Nelson, D. y Cox, M. (2004). *Lehninger Principles of Biochemistry*. W. H. Freeman, cuarta edición. 1100 pp.

Pabo, C. (1983). Molecular technology: Designing proteins and peptides. *Nature*, **301**: 200.

Palsson, B. (2000). The challenges of in silico biology. *Nature biotechnology*, **18**(11): 1147.

Pantazes, R., Grisewood, M., y Maranas, C. (2011). Recent advances in computational protein design. *Current Opinion in Structural Biology*, **21**(4): 467–472.

Park, B. y Levitt, M. (1996). Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *Journal of molecular biology*, **258**(2): 367–392.

Parsons, J., Holmes, J., Rojas, J., Tsai, J., y Strauss, C. (2005). Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of computational chemistry*, **26**(10): 1063–1068.

Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T., DeBolt, S., Ferguson, D., Seibel, G., y Kollman, P. (1995). Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, **91**(1): 1–41.

Peterson, L., Kang, X., y Kihara, D. (2014). Assessment of protein side-chain conformation prediction methods in different residue environments. *Proteins: Structure, Function, and Bioinformatics*, **82**(9): 1971–1984.

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

Petrella, R., Lazaridis, T., y Karplus, M. (1998). Protein sidechain conformer prediction: a test of the energy function. *Folding and Design*, **3**(5): 353–377.

Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L., y Schulten, K. (2005). Scalable molecular dynamics with namd. *Journal of computational chemistry*, **26**(16): 1781–1802.

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.

Quan, L., Lü, Q., Li, H., Xia, X., y Wu, H. (2014). Improved packing of protein side chains with parallel ant colonies. *BMC bioinformatics*, **15**(Suppl 12): S5.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B., Das, R., Grishin, N., y Baker, D. (2009). Structure prediction for casp8 with all-atom refinement using rosetta. *Proteins: Structure, Function, and Bioinformatics*, **77**(S9): 89–99.

Rapaport, D. (2004). *Art of molecular dynamics simulation*. Cambridge Univ. Press.

Richards, F. (1977). Areas, Volumes, Packing, and Protein Structure. *Annual Review of Biophysics and Bioengineering*, **6**: 151–176.

Roberts, K., Cushing, P., Boisguerin, P., Madden, D., y Donald, B. (2012). Computational design of a pdz domain peptide inhibitor that rescues cftr activity. *PLoS computational biology*, **8**(4): e1002477.

Rodríguez, D. (2014). *Diseno de heuristicas para el problema de empaquetamiento de la cadena lateral en proteinas*. Tesis de maestría, CICESE.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T., y Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, **14**(6): 523–528.

Rohl, C., Strauss, C., Misura, K., y Baker, D. (2004). Protein structure prediction using rosetta. *Numerical Computer Methods, Part D, Methods in Enzymology*, **383**: 66–93.

Roy, A., Kucukural, A., y Zhang, Y. (2010). I-tasser: a unified platform for automated protein structure and function prediction. *Nature protocols*, **5**(4): 725.

Ryu, J., Lee, M., Cha, J., Laskowski, R. A., Ryu, S. E., y Kim, D.-S. (2016). Betascpweb: side-chain prediction for protein structures using voronoi diagrams and geometry prioritization. *Nucleic acids research*, **44**: W416—-W423.

Samudrala, R. y Moult, J. (1998). Determinants of side chain conformational preferences in protein structures. *Protein Engineering*, **11**(11): 991–997.

Sasisekharan, V. y Ponnuswamy, P. (1970). Backbone and side-chain conformations of amino acids and amino acid residues in peptides. *Biopolymers*, **9**(10): 37–45.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8. Recuperado (31/05/2018) de: https://pymol.org/2/.

Shapovalov, M. y Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**(6): 844–858.

Shen, M. y SaliProtein, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, **15**(11): 2507–2524.

Simons, K. T., Bonneau, R., Ruczinski, I., y Baker, D. (1999). Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins: Structure, Function, and Bioinformatics*, **37**(S3): 171–176.

Smadbeck, J., Bellows, M., Khoury, G., Taylor, M., y Floudas, C. (2013). Protein WISDOM: A Workbench for In silico De novo Design of BioMolecules. *Journal of Visualized Experiments*, **77**: 1–25.

Smolinski, T., Milanova, M., y Hassanien, A. (2009). *Computational Intelligence in Biomedicine and Bioinformatics: Current trends and applications*, Vol. 151.

Soto, C. (2001). Protein misfolding and disease; protein refolding and therapy. *FEBS letters*, **498**(2-3): 204–207.

Stelling, J., Kremling, A., Ginkel, M., Bettenbrock, K., y Gilles, E. (2001). *Foundations of Systems Biology*. MIT press.

Suarez, M. y Jaramillo, A. (2009). Challenges in the computational design of proteins. *Journal of the Royal Society, Interface*, **6**(4): S477–S491.

Summers, N. y Karplus, M. (1989). Construction of side-chains in homology modelling: Application to the C-terminal lobe of rhizopuspepsin. *Journal of Molecular Biology*, **210**(4): 785–811.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, **1**(2): 146–160.

Touw, W., Baakman, C., Black, J., te Beek, T., Krieger, E., Joosten, R., y Vriend, G. (2014). A series of pdb-related databanks for everyday needs. *Nucleic acids research*, **43**(D1): D364–D368.

Turanli-Yildiz, B., Alkim, C., y Cakar, Z. (2012). Protein engineering methods and applications. En: *Protein Engineering*. InTech.

van den Berg, B., Reinders, M., van der Laan, J., Roubos, J., y de Ridder, D. (2014). Protein redesign by learning from data. *Protein Engineering, Design & Selection*, **27**(9): 281–288.

Vendruscolo, M. y Dobson, C. (2011). Protein dynamics: Moore's law in molecular biology. *Current biology*, **21**(2): R68–R70.

Vlieghe, P., Lisowski, V., Martinez, J., y Khrestchatisky, M. (2010). Synthetic therapeutic peptides: science and market. *Drug discovery today*, **15**(1): 40–56.

Voigt, C., Gordon, D., y Mayo, S. (2000). Trading Accuracy for Speed: A Quantitative Comparison of Search Algorithms in Protein Sequence Design. *Journal of Molecular Biology*, **299**(3): 789–809.

Wang, G. y Dunbrack, R. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, **19**(12): 1589–1591.

Wendoloski, J. y Salemme, F. (1992). PROBIT: A statistical approach to modeling proteins from partial coordinate data using substructure libraries. *Journal of Molecular Graphics*, **10**(2): 124–126.

Wijma, H. y Janssen, D. (2013). Computational design gains momentum in enzyme catalysis engineering. *The FEBS journal*, **280**(13): 2948–2960.

Wilson, C., Gregoret, L., y Agard, D. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *Journal of molecular biology*, **229**(4): 996–1006.

Wlodawer, A., Minor, W., Dauter, Z., y Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal*, **275**(1): 1–21.

Woolfson, D., Bartlett, G., Burton, A., Heal, J., Niitsu, A., Thomson, A., y Wood, C. (2015). De novo protein design: how do we expand into the universe of possible protein structures? *Current opinion in structural biology*, **33**: 16–26.

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

Xu, D. y Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, **80**(7): 1715–1735.

Zhang, Y. (2008). I-tasser server for protein 3d structure prediction. *BMC bioinformatics*, **9**(1): 40.

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

Anexo

A.1. Obtención de las coordenadas de los átomos de la cadena lateral en función a los ángulos de torsión y parámetros de longitudes y ángulos entre enlaces

Los datos en las bibliotecas de rotámeros usualmente son los ángulos de torsión de las conformaciones representadas, mientras que ciertos términos de funciones de score necesitan las coordenadas cartesianas de los átomos de la cadena lateral (el término de Van der Waals por ejemplo). Por lo tanto, es necesario contar con un método rápido y preciso para realizar la conversión.

En la Subsección 2.4.1 se indicó que se puede obtener todas las coordenadas de los átomos de la cadena lateral de un residuo a partir de las coordenadas de los átomos de la cadena principal y de los ángulos de torsión de la cadena lateral. Para ello, se debe contar con valores estándar de longitudes de enlaces covalentes y de ángulos entre estos enlaces. También se debe tener en cuenta que existen átomos *principales* y átomos *redundantes* en la cadena lateral, dependiendo del tipo de residuo considerado.

A.1.1. Problema geométrico: de ángulos de torsión a coordenadas 3D

Como primer paso, es necesario contar con un método para obtener las coordenadas tridimensionales $\vec{D}$ de un punto D (i.e., $\vec{D} \in \mathbb{R}^3$), a partir de los siguientes datos:

- Los vectores de posición $\vec{A}, \vec{B}, \vec{C} \in \mathbb{R}^3$ de los puntos de referencia.
- La longitud L del vector $\overrightarrow{CD}$.
- El ángulo θ entre los vectores $\overrightarrow{CD}$ y $\overrightarrow{CB}$.
- El ángulo de torsión o ángulo diedro φ entre los planos α (determinado por $\vec{A}$, $\vec{B}$ y $\vec{C}$) y β (determinado por $\vec{B}$, $\vec{C}$ y $\vec{D}$).

Considerando la representación de la estructura de una proteína, A, B, C, D son los átomos y $\vec{A}, \vec{B}, \vec{C}, \vec{D} \in \mathbb{R}^3$ son las respectivas posiciones; L es la longitud del enlace

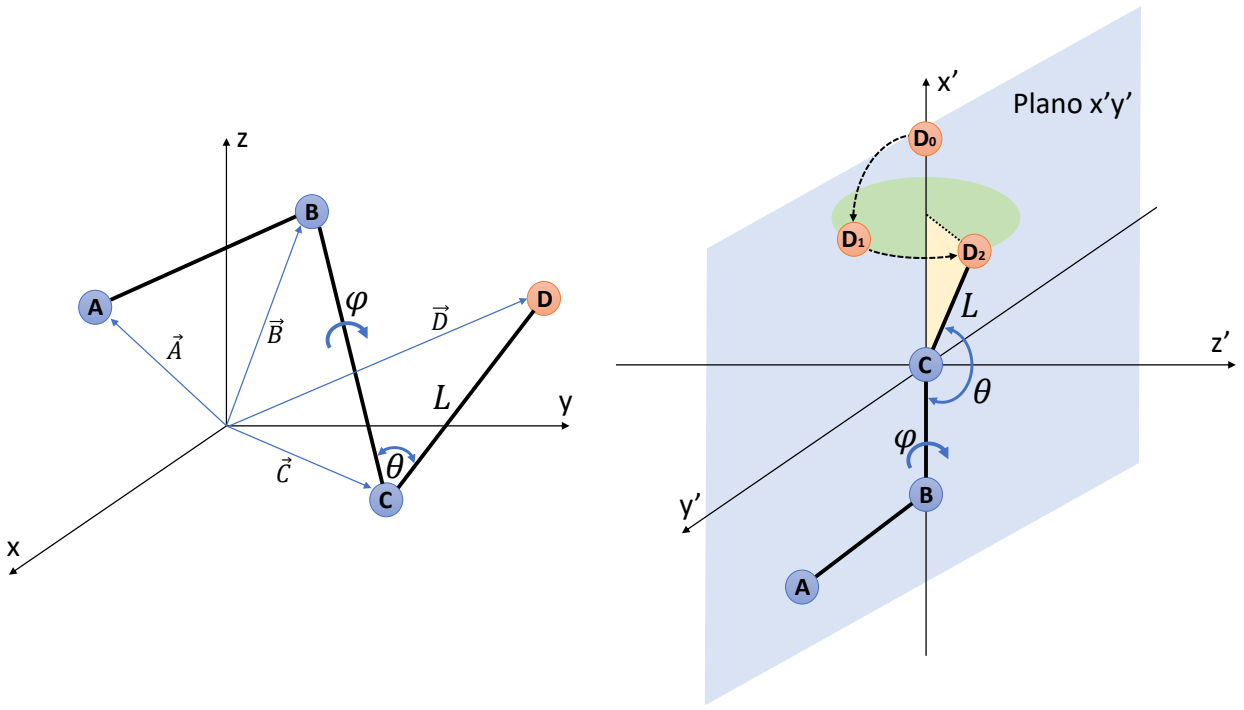

Figura 38. Problema geométrico: de ángulos de torsión a coordenadas 3D. En la parte izquierda se muestran los datos de entrada, siendo la salida el vector $\vec{D}$. La parte derecha es a una adaptación de la representación de la propuesta de Parsons *et al.* (2005). El punto *C* es el origen del sistema auxiliar $x'y'z'$, el punto *B* se encuentra en el eje x' negativo, mientras que el punto *A* se encuentra en el plano $x'y'$. El punto D_0 resulta de posicionarse en el eje x' positivo a una distancia *L* de *C*. El punto D_1 resulta de girar D_0 dentro de una circunferencia con centro en *C* y radio *L* dentro del plano $x'y'$, de tal forma que *BC* y CD_1 formen un ángulo θ. Por último, el punto D_2 resulta de girar D_1 alrededor del eje x', de tal forma que el plano determinado por *B*, *C* y D_2 forme un ángulo φ con el plano $x'y'$ (determinado por *A*, *B* y *C*). A partir de D_2 se obtiene *D* mediante un cambio del sistema de referencia, lo cual se realiza con la matriz *M*.

covalente entre *C* y *D*, θ es el ángulo entre los enlaces *CD* y *BC*, mientras que φ es el ángulo de torsión respecto al enlace *BC*.

Parsons *et al.* (2005) proponen el método *NeRF* (*Natural extension Reference Frame*), que mejora a otros existentes para realizar la conversión de ángulos de torsión a coordenadas tridimensionales. Este método es tanto rápido como fácil de implementar, y es el empleado por Rosetta (Rohl *et al.*, 2004) para el manejo de estructuras de proteínas.

El método *NeRF* emplea un marco de referencia auxiliar x', y', z' donde el átomo *C* está en el origen de este marco, el átomo *B* está en el eje x' negativo y el átomo *A* se encuentra en el plano $x'y'$. Entonces, la posición inicial del átomo *D* bajo este marco (representado por el vector de posición $\vec{D}_2$) está dada por la transformación usual del

sistema de coordenadas esféricas al de coordenadas cartesianas:

$$\vec{D}_2 = \left(L \times \cos(\theta),\ L \times \cos(\varphi)\,\mathrm{sen}(\theta),\ L \times \sin(\varphi)\,\mathrm{sen}(\theta)\right) \tag{20}$$

Por lo tanto, $\vec{D}_2$ no depende de las posiciones de A, B y C; sino de los valores L, θ y φ. Para obtener el vector de posición $\vec{D}$ del átomo D, se pasa del marco auxiliar al utilizado para representar a los vectores de posición $\vec{A}$, $\vec{B}$ y $\vec{C}$. Para ello, se siguen estos pasos:

- Se determina el vector $\overrightarrow{BC}$ y se normaliza el mismo, obteniéndose $\overrightarrow{bc}$.
- Se calcula el vector $\vec{n}$, que resulta de la normalización del producto vectorial entre los vectores $\overrightarrow{AB}$ y $\overrightarrow{bc}$.
- Se obtiene la matriz de cambio de referencia M, donde las columnas están dadas por: (i) el vector $\overrightarrow{bc}$, (ii) el producto vectorial entre $\vec{n}$ y $\overrightarrow{bc}$, y (iii) el vector $\vec{n}$.
- Finalmente: $\vec{D} = M\vec{D}_2 + \vec{C}$.

De esta manera, para generar todos los átomos de la cadena principal de una cadena polipeptídica a partir de los ángulos de torsión correspondientes a cada residuo, sólo basta conocer las posiciones de los tres átomos de la cadena principal del primer residuo; realizándose el procedimiento descrito para obtener la posición de cada átomo de manera secuencial en función a los determinados anteriormente, hasta llegar al final de la cadena. Esto puede verse en la Figura 39. De manera análoga, también puede determinarse las posiciones de cada átomo de la cadena lateral a partir de átomos de referencia de la cadena principal del residuo y los ángulos de torsión de la cadena lateral. Esto se verá con más detalle a continuación, y en la Figura 39 se muestra un ejemplo con la metionina (MET).

A.1.2. Parámetros para la construcción de residuos a partir de ángulos de torsión

Como en el PSCPP ya se conoce la ubicación de cada átomo de la cadena principal de la proteína, la atención se centra en la generación de los átomos de la cadena lateral de cada residuo a partir de los ángulos de torsión respectivos. El valor de la

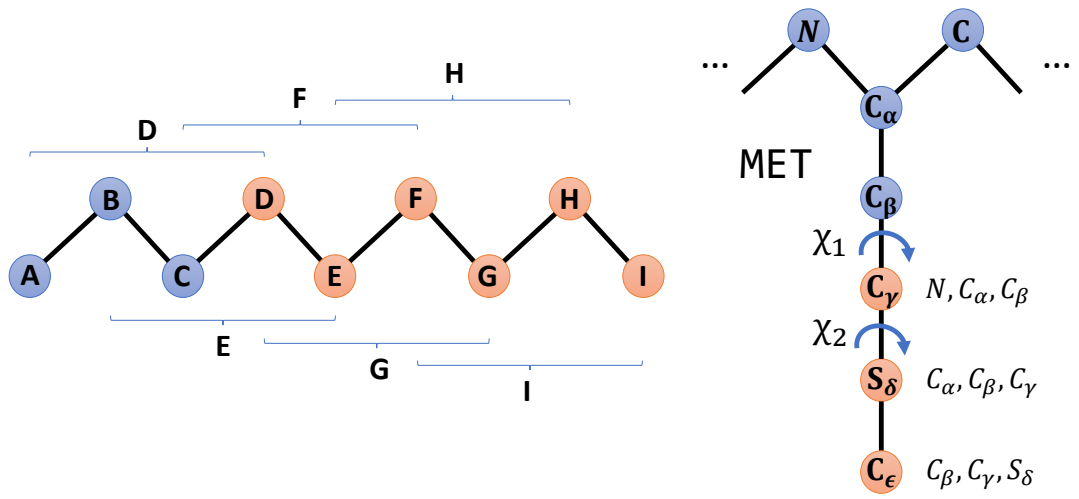

Figura 39. Cálculo secuencial de las coordenadas 3D de los átomos de la cadena lateral. En la parte izquierda se muestra el proceso secuencial para la obtención de las coordenadas de todos los átomos de una cadena polipeptídica. En la parte derecha se muestra el proceso de obtención de los átomos de la cadena lateral de la metionina (MET).

longitud del enlace L y el valor del ángulo entre enlaces θ dependen de los tipos de átomos que forman estos enlaces, y están restringidos debido a la naturaleza de los enlaces covalentes. Por lo tanto, estos valores pueden considerarse independientes de la estructura de la proteína considerada. Los valores de Engh y Huber (1991) fueron seleccionados en el presente trabajo ya que son comúnmente empleados en métodos del estado del arte para el PSCPP (Liang y Grishin, 2002; Liang *et al.*, 2011a; Miao *et al.*, 2011), y se obtuvieron a partir de un análisis estadístico de estructuras almacenadas en el PDB. En la Tabla 21 se muestra el tipo de átomo correspondiente a cada uno de los átomos de cada tipo de residuo (de los 20 naturales), de acuerdo al trabajo de Engh y Huber (1991).

Por otro lado, puede observarse en la Tabla 1 que no todos los átomos sirven de referencia para obtener los ángulos de torsión de la cadena lateral. Por ejemplo, la fenilalanina (PHE) tiene sólo dos ángulos de torsión (χ_1 y χ_2); pero tiene más de dos átomos en su cadena lateral (Ver Figura 6). Esto se debe a que la estructura de anillo en su cadena lateral restringe las posiciones que pueden tomar los átomos que la componen. Por lo tanto, conviene hacer una clasificación de los átomos de la cadena lateral de un residuo: los átomos *principales* de la cadena lateral son los átomos de la cadena lateral necesarios para calcular los ángulos de torsión de la cadena lateral; y los demás se conocen como átomos *redundantes* de la cadena lateral. Los átomos redundantes pueden obtenerse a partir de la ubicación de los átomos principales. Por

ejemplo, el átomo CG2 de la valina (VAL) puede obtenerse con los mismos átomos de referencia que CG1, solamente debe sumarse 120° al ángulo de torsión χ_1 empleado para determinar CG1.

Un caso particular constituye el átomo de carbono beta (C_β). Este átomo está en función a los átomos de referencia N, C_α y C. Si bien no existe un enlace entre C y C_β, estadísticamente se ha encontrado una relación con el ángulo de torsión $N{-}C_\alpha{-}C{-}C_\beta$ para cada tipo de residuo. Por esta razón, en términos prácticos este átomo también se considera como entrada para un caso del PSCPP, por lo que su posición no es predicha por un método en particular. Ciertos métodos como SCWRL4 (Krivov *et al.*, 2009), RASP (Miao *et al.*, 2011), CIS-RR (Cao *et al.*, 2011) y SCCOMP (Eyal *et al.*, 2004); agregan el átomo C_β en caso de que no se encuentre en la estructura de entrada. Para nuestros experimentos, también se agrega C_β de acuerdo a los datos en SCCOMP.

A.1.2.1. Presentación del archivo con los datos para la construcción de residuos, además de los radios de Van der Waals para cálculo de colisiones

En el enlace `https://figshare.com/s/e0a29a17de318223bc26` puede descargarse el archivo `Anexo.zip`, que contiene el archivo `parametrosEnghYHuber.txt`. Este último contiene los valores de longitudes de enlace, ángulos entre enlaces y ángulos de torsión (para C_β y átomos redundantes de la cadena lateral); además de los átomos de referencia necesarios para determinar cada átomo de la cadena lateral. Este archivo también contiene los radios de Van der Waals para cada átomo de cada tipo de residuo, y estos radios sirven para determinar la cantidad de colisiones en una determinada estructura de acuerdo al criterio establecido en la Subsección 3.6.4. En la figura 40 se muestra la presentación de los datos correspondientes a la valina (VAL).

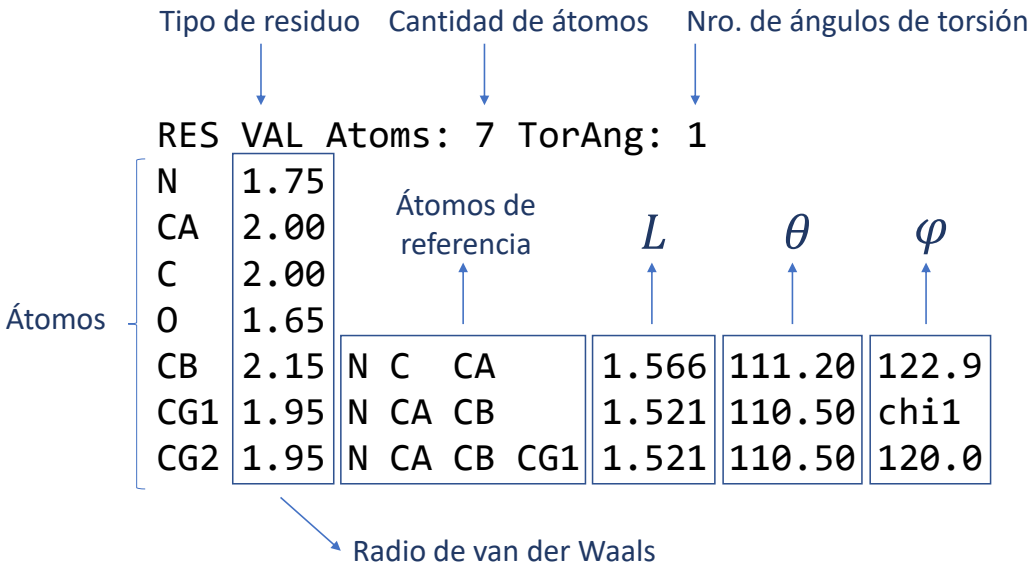

Figura 40. Presentación del archivo de parámetros, mostrando la valina (VAL) como ejemplo. En la primera línea se indica el tipo de residuo, la cantidad total de átomos y la cantidad de ángulos de torsión en la cadena lateral. Para los átomos de la cadena principal sólo se indican los correspondientes radios de Van der Waals de acuerdo al trabajo de Cao *et al.* (2011). Cada línea de los átomos restantes contiene además los datos para obtener el punto *D* a partir de los átomos de referencia *A*, *B* y *C* (ver Figura 38), los cuales son: (i) las posiciones de los átomos de referencia *A*, *B* y *C*; (ii) la longitud *L* del enlace *C*-*D*, (iii) el ángulo θ entre los enlaces *CB* y *CD*, y (iv) el ángulo de torsión φ entre los planos $A-B-C$ y $B-C-D$. Si en lugar de una constante se tiene "chi_i" ($i \in \{1, 2, 3, 4, 5\}$), entonces este ángulo de torsión depende del rotámero empleado para dicho residuo. Si aparece un cuarto átomo de referencia *E*, esto implica que el átomo *D* es un átomo redundante de la cadena lateral; por ello, debe calcularse el ángulo de torsión entre $A-B-C$ y $B-C-E$ y sumarle el valor de la última columna para obtener φ. Para la valina, este caso se da para el cálculo de CG2.

Tabla 21. Átomos para cada residuo y los tipos correspondientes en el trabajo de Engh y Huber (1991).

Residuo	Cadena principal				Cadena lateral											
					Átomos principales					Átomos secundarios						
ALA	N NH1	CA CH1E	C C	O O	CB CH3E											
ARG	N NH1	CA CH1E	C C	O O	CB CH2E	CG CH2E	CD CH2E	NE NH1	CZ C	NH1 NC2	NH2 NC2					
ASN	N NH1	CA CH1E	C C	O O	CB CH2E	CG C	OD1 O			ND2 NH2						
ASP	N NH1	CA CH1E	C C	O O	CB CH2E	CG C	OD1 OC			OD2 OC						
CYS	N NH1	CA CH1E	C C	O O	CB CH2E	SG SH1E										
GLN	N NH1	CA CH1E	C C	O O	CB CH2E	CG CH2E	CD C	OE1 O		NE2 NH2						
GLU	N NH1	CA CH1E	C C	O O	CB CH2E	CG CH2E	CD C	OE1 OC		OE2 OC						
GLY	N NH1	CA CH2G	C C	O O												
HIS	N NH1	CA CH1E	C O	O C	CB CH2E	CG C5	ND1 NR			CD2 CR1H	CE1 CRHH	NE2 NH1				
ILE	N NH1	CA CH1E	C O	O C	CB CH1E	CG1 CH2E	CD1 CH3E			CG2 CH3E						
LEU	N NH1	CA CH1E	C O	O C	CB CH2E	CG CH1E	CD1 CH3E			CD2 CH3E						
LYS	N NH1	CA CH1E	C O	O C	CB CH2E	CG CH2E	CD CH2E	CE CH2E	NZ NH3							
MET	N NH1	CA CH1E	C O	O C	CB CH2E	CG CH2E	SD SM	CE CH3E								
PHE	N NH1	CA CH1E	C O	O C	CB CH2E	CG CF	CD1 CR1E			CD2 CR1E	CE1 CR1E	CE2 CR1E	CZ CR1E			
PRO	N N	CA CH1E	C O	O C	CB CH2E	CG CH2P	CD CH2P									
SER	N NH1	CA CH1E	C O	O C	CB CH2E	OG OH1G										
THR	N NH1	CA CH1E	C O	O C	CB CH1E	OG1 OH1				CG2 CH3E						
TRP	N NH1	CA CH1E	C O	O C	CB CH2E	CG C5W	CD1 CR1E			CD2 CW	NE1 NH1	CE2 CW	CE3 CR1E	CZ2 CR1W	CZ3 CR1E	CH2 CR1W
TYR	N NH1	CA CH1E	C O	O C	CB CH2E	CG CY	CD1 CR1E			CD2 CR1E	CE1 CR1E	CE2 CR1E	CZ CY2	OH OH1		
VAL	N NH1	CA CH1E	C O	O C	CB CH1E	CG1 CH3E				CG2 CH3E						

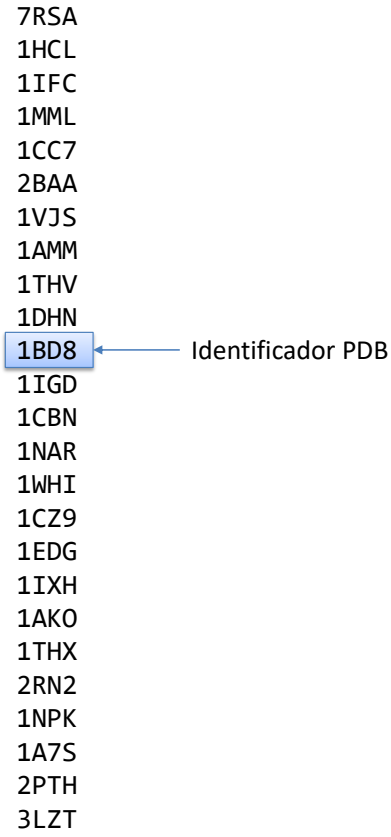

Figura 41. Presentación del archivo con los identificadores PDB para cada conjunto de prueba. Como ejemplo, se muestra el caso del conjunto de 25 proteínas, empleado en las simulaciones de dinámica molecular.

A.2. Listas de conjuntos de prueba de proteínas

En el archivo `Anexo.zip` en `https://figshare.com/s/e0a29a17de318223bc26`, se encuentra la carpeta `listas`. Aquí se encuentran los identificadores PDB para cada proteína en cada conjunto de prueba empleado en el presente trabajo de investigación. En `https://www.rcsb.org/pages/download_features#Structures` puede colocarse el contenido del archivo para un conjunto de prueba en particular, de manera a descargar todos los archivos de una sola vez. Como ejemplo, en la Figura 41 se muestra el contenido del archivo `lista25.txt`, que corresponde al conjunto de prueba de 25 proteínas.

A.3. Parámetros de las funciones de score implementadas

En la Sección 4.1 se describió cada uno de los términos de las dos funciones de score implementadas. Ciertos parámetros no fueron mostrados en el documento principal. Por ello, en el archivo `Anexo.zip` en `https://figshare.com/s/e0a29a17de318223bc26`, se encuentra una carpeta `paramsFS` que contiene los parámetros faltantes para imple-

mentar las funciones de score de CIS-RR (Cao *et al.*, 2011) y RASP (Miao *et al.*, 2011).

En el archivo `CISRR.txt` están los radios de Van der Waals para cada átomo de cada uno de los 20 residuos naturales.

En el archivo `RASP.txt` están, además de los radios de Van der Waals, las profundidades para cada átomo de cada uno de los 20 residuos naturales. También contiene los pesos para cada residuo para el cálculo del término relativo a las frecuencias de los rotámeros en la biblioteca.

A.4. Criterio para la determinación de vecindad entre residuos

En la Subsección 4.1.3 se estableció que son pre-calculadas solamente las interacciones entre átomos vecinos. Para determinar la lista de vecinos para cada residuo, para cada tipo de residuo se calculó la máxima distancia de sus átomos al C_α (carbono-alfa), considerando cada rotámero en la biblioteca de Dunbrack y Cohen (1997) para ese tipo de residuo. Dentro del archivo `Anexo.zip` en `https://figshare.com/s/e0a29a17de318223bc26`, se encuentra el archivo `maxDistACAlfa.txt` con los resultados para cada uno de los 20 residuos naturales, indicando además cuál es el átomo a mayor distancia y la conformación de la cadena lateral necesaria para que aparezca dicha distancia. Con esta información, dos residuos son considerados vecinos si la distancia entre sus átomos C_α es menor que la suma de sus máximas distancias a C_α, más dos veces el máximo radio de Van der Waals para cada uno de los residuos considerados. Esta condición se ilustra con la ayuda de la Figura 42.

A.5. Aspectos relacionados a las simulaciones de dinámica molecular

En la Sección 5.2 se dio una descripción general de las simulaciones de dinámica molecular para las estructuras cristalográficas de referencia, y en esta sección se brindan mayores detalles. Se usó VMD (Humphrey *et al.*, 1996) para agregar los átomos de hidrógeno a las estructuras experimentales en el conjunto de prueba de 25 proteínas. Cada estructura experimental fue puesta en una caja llena de agua, cuya dimensión fue definida de tal forma que exista una capa de agua de 15 Å en cada dirección a partir del átomo con la coordenada más alta (en valor absoluto) en dicha dirección. El sis-

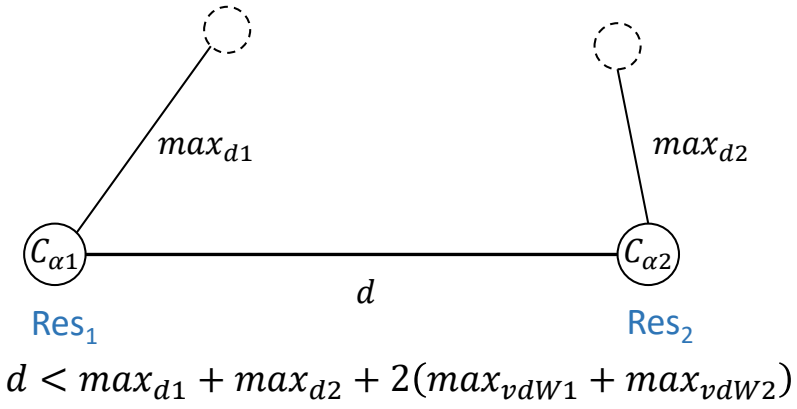

Figura 42. Condición de vecindad entre dos residuos en función a las máximas distancias a C_α y los radios de Van der Waals.

tema fue neutralizado añadiendo Na^+ y Cl^-. Los parámetros (par_all27_prot_lipid.prm) y la función de energía de CHARMM (Brooks *et al.*, 2009) fueron empleados en NAMD (Phillips *et al.*, 2005) para la minimización de energía, el calentamiento y la dinámica molecular por recocido simulado (MD-SA).

El proceso de minimización se realizó empleando el gradiente conjugado y el algoritmo de búsqueda lineal por omisión en NAMD con 1000 pasos. Luego, el resultado pasó a un proceso de calentamiento sin restricciones, con volumen constante y temperatura modificada por el método de recocido simulado, compuesto de cuatro etapas:

- Etapa 1: un proceso de calentamiento por 0.5 ns, aumentando gradualmente la temperatura desde 50 K hasta 400 K.
- Etapa 2: se mantuvo la temperatura a 400 K por 0.5 ns.
- Etapa 3: se disminuyó la temperatura gradualmente desde 400 K hasta 310 K en 0.2 ns.
- Etapa 4: se mantuvo la temperatura a 310 K por 0.8 ns.

El último paso de todo el proceso fue la simulación principal de dinámica molecular a presión y temperatura constante (NPT) por 50 ns, a 310 K, con conformaciones guardadas cada 20 ps. Se aplicó condiciones periódicas de contorno. Los enlaces que involucran a moléculas de agua fueron restringidos, y se empleó pasos de 2 fs en las simulaciones de dinámica molecular. La temperatura fue controlada con un termostato de Langevin (Loncharich *et al.*, 1992).

En el archivo `Anexo.zip` en `https://figshare.com/s/e0a29a17de318223bc26` se encuentra la carpeta `dinamicaMolecular` con todos los archivos de configuración en NAMD necesarios para la minimización, el calentamiento y el proceso principal de dinámica molecular.

A.6. Diseño de funciones de score

Luego de evaluar el desempeño de las funciones de score de algunos métodos del estado del arte, el siguiente paso es diseñar mejores funciones de score. Un camino para ello es diseñar una función de score compuesta de los dos términos más importantes: el que considera las interacciones de Van der Waals (término de Van der Waals) y el término en función a las conformaciones de los rotámeros en la biblioteca (término de ángulos de torsión). Para el diseño de estos términos podría emplearse programación genética (Koza, 1994), o bien cada término podría tener una forma predefinida. Este segundo enfoque fue el elegido para las pruebas preliminares, basándose en el trabajo de Liang *et al.* (2011b). El término de Van der Waals se define por:

$$E_{(d)} = a_1 \times d^{-2} + a_2 \times d^{-4} + a_3 \times d^{-6} + a_4 \times d^{-8} \tag{21}$$

donde d es la distancia entre los átomos y a_{1-4} son parámetros que dependen del par de átomos involucrados (se consideró 14 tipos de átomos). El término de ángulos de torsión está dado por:

$$E_{torsion} = t_1 \times \cos\alpha + t_2 \times \sin\alpha + t_3 \times \cos 2\alpha + t_4 \times \sin 2\alpha + t_5 \times \cos 3\alpha + t_6 \times \sin 3\alpha \tag{22}$$

donde α es el ángulo de torsión de la cadena lateral del rotámero y t_{1-6} son parámetros que corresponden a un ángulo de torsión particular de cierto tipo de residuo. Existen 39 ángulos de torsión para los 20 residuos naturales.

Por lo tanto, existe un total de 654 parámetros a ser definidos. La principal diferencia entre esta propuesta y el trabajo de Liang *et al.* (2011b) es que, para la optimización de estos parámetros, se plantea usar el método de búsqueda local propuesto en el presente trabajo para medir la precisión al final de la búsqueda local (comenzando con la estructura de referencia) para un conjunto de proteínas. Se conjetura que esta evaluación proveerá mayor información que permitirá distinguir de mejor manera la

Tabla 22. Descripción del algoritmo genético (GA).

Representación	Arreglo de números reales
Recombinación	Aritmética ($\alpha = 0.5$)
Probabilidad de Recombinación	100 %
Mutación	Perturbación Gaussiana
P. Gaussiana - Desviación estándar	0.01
Probabilidad de Mutación	100 %
Selección de padres	Torneo binario
Selección de sobrevivientes	Generacional
Porcentaje de Elitismo	10 %
Tamaño de la Población	120
Número de generaciones	2000
Número de hijos	120
Inicialización	Aleatoria

calidad de dos funciones de score distintas.

A.6.1. Pruebas preliminares

Como se mencionó, bajo este esquema de diseño se tiene que definir 654 parámetros; por lo que una solución está dada por un arreglo de 654 números reales. Para las pruebas preliminares, cada solución candidata se evaluó empleando el conjunto de prueba de 25 proteínas. Inicialmente se utilizó un algoritmo genético (GA por sus siglas en inglés), aunque posteriormente se cambió a una estrategia evolutiva (ES por sus siglas en inglés) (Eiben *et al.*, 2003). Esto fue debido a que las mutaciones por perturbación con desviación estándar constante no lograban mejorar los resultados a partir de una cierta cantidad de iteraciones y se llegaba a una convergencia prematura en el GA. En las tablas 22 y 23 se muestran los parámetros empleados para cada método.

Un punto importante es que si sólo consideramos la estructura de referencia como entrada al algoritmo de búsqueda local, puede darse el caso de que tengamos un desempeño óptimo bajo el método de búsqueda local si todos los parámetros de la función de score son iguales a cero; pues en este caso se tendrá un score igual a cero para cualquier posible solución. Por ello, se pondera el desempeño de la búsqueda local comenzando con *Native*, *MostProb* y *Random*[1]; otorgándoles a sus resultados de precisión distintos pesos. Como se tienen dos medidas de precisión (χ_1 y χ_{1+2}), cada una tendrá también un peso para determinar la calidad de una cierta función de score

[1]Ver Sección 4.2 para las definiciones de estas estructuras.

Tabla 23. Descripción de la estrategia evolutiva (ES).

Representación	Arreglo de números reales
Recombinación	Aritmética ($\alpha = 0.5$)
Probabilidad de Recombinación	100 %
Mutación	Perturbación Gaussiana
P. Gaussiana - Desviación estándar inicial	0.01
P. Gaussiana - Mínima desviación estándar	0.000001
Tipo de Mutación	No correlacionada, con 654 tamaños de salto
Selección de padres	Aleatoria uniforme
Selección de sobrevivientes	(μ, λ)
Porcentaje de Elitismo	10 %
Tamaño de la Población	20
Número de generaciones	2000
Número de hijos	120
Inicialización	Aleatoria

Tabla 24. Otros parámetros para las pruebas experimentales. Si sólo se considera la estructura de referencia, puede darse el caso de que se tenga un desempeño óptimo bajo el método de búsqueda local si todos los parámetros de la función de score son iguales a cero. Por ello, se pondera el desempeño de la búsqueda local comenzando con *Native*, *MostProb* y *Random*. Entre paréntesis se muestran los pesos empleados para cada componente.

Valores - Límite inferior	-1.0
Valores - Límite superior	1.0
Estructuras iniciales	*Native* (0.7), *MostProb* (0.2) y *Random* (0.1)
Medidas de precisión	χ_1 (0.8) y χ_{1+2} (0.2)

candidata. También se estableció límites para los valores de cada parámetro, pues lo más importante es la relación entre ellos; y con esto también se evita tener valores absolutos de score muy grandes. Los resultados de las pruebas experimentales se muestran en la Tabla 25.

Las precisiones alcanzadas por la funciones de score diseñadas están aún bastante alejadas de las obtenidas por los métodos del estado del arte. Inclusive, al usar el algoritmo de búsqueda local con las funciones de score de CIS-RR y RASP se logran resultados considerablemente mejores. Además, la cantidad de colisiones que aparecen en las funciones Liang (AG) y Liang (ES) es muy alta. Todo esto podría indicar que no debería considerarse únicamente la precisión como métrica de calidad a la hora de evaluar funciones de score candidatas, sino también considerar el número de colisiones, la variación de energía, etc.

Tabla 25. Valores de referencia para la precisión de las estructuras en el conjunto de prueba de 25 proteínas, y resultados de las pruebas experimentales del diseño de funciones de score mediante el método de Liang *et al.* (2011b). Para los valores de referencia, se muestran los resultados considerando o no los contactos simétricos en la estructura experimental (WHATIF y Experimental, respectivamente).

Resultados de referencia
Precisión total y número de colisiones

	Experimental			WHATIF		
	χ_1(%)	χ_{1+2}(%)	Colisiones	χ_1(%)	χ_{1+2}(%)	Colisiones
SCWRL4	83.73	73.6	29	84.47	75.04	32
OPUS-Rota	85.76	75.92	48	86.63	77.4	51
CIS-RR	84.47	73.83	6	85.34	75.66	7
RASP	83.91	72.33	46	84.82	73.14	75
SIDEpro	84.92	74.84	5	86.3	76.84	4

Resultados de las pruebas experimentales
Precisión total y número de colisiones

	Native			*MostProb*			*Random*		
	χ_1(%)	χ_{1+2}(%)	Colisiones	χ_1(%)	χ_{1+2}(%)	Colisiones	χ_1(%)	χ_{1+2}(%)	Colisiones
Liang (AG)	66.95	41.87	4252	61.04	32.84	4712	62.69	34.97	4728
Liang (ES)	84.87	68.73	850	72.86	53.06	1201	71.97	52.18	1278
CIS-RR	89.44	82.99	2	83.51	71.84	36	83.51	71.84	56
RASP	89.52	84.1	4	83.39	71.93	50	82.97	71.15	78