UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

PREDICCIÓN DE LA ESTRUCTURA TRIDIMENSIONAL DE LAS PROTEÍNAS A PARTIR DE SU SECUENCIA

TESIS

QUE PARA OBTENER EL TíTULO DE:

Licenciado en Física

PRESENTA:

Víctor Martinell García

DIRECTOR DEL TRABAJO:

Gabriel del Río Guerra

2016

Ciudad Universitaria, CDMX

1. Datos del alumno
Martinell
García
Víctor
21 56 23 24
Universidad Nacional Autónoma de México
Facultad de Ciencias
Física
307511861

2. Datos del asesor
Doctor
del Río
Guerra
Gabriel

3. Datos del sinodal 1
Doctor
Villarreal
Luján
Carlos

4. Datos del sinodal 1
Doctor
Cordero
Barboa
Adolfo Ernesto

5. Datos del sinodal 1
Doctor
Benet
Fernández
Luis

6. Datos del sinodal 1
Doctor
Miramontes
Vidal
Pedro Eduardo

7. Datos de la tesis
Predicción de la Estructura Tridimensional
de las Proteínas a partir de su Secuencia
72 p.
2016

A mis padres Julio y Mercedes; por todo el tiempo que me han aguantado y todo el que todavía me tienen que aguantar.

Agradecimientos

Muchas gracias a Paula y Diana, por entenderme con tanta facilidad, y siempre dedicarme el mayor tiempo posible. También a mis papás por todo lo que han hecho por mi y hacerme gran parte de lo que soy.

Quiero agradecer a Gabriel y a todos miembros del laboratorio, en especial a Ricardo, por hacer mi estancia tan amena e ilustrativa.

Gracias a mis amigos Joana, Francisco Cynthia y los otros raros. También a la gente de cómputo, por acompañarme en mi segunda carrera.

Gracias a mi Maestro, Francisco, que, sin que me diera cuenta, me ha enseñado tantísimas cosas.

Finalmente quiero agradecer a la Dra. María Teresa Lara Ortiz por su apoyo técnico y al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (IN208014) el cual financió parcialmente este trabajo.

Resumen

En este trabajo se intenta resolver el problema de predicción del plegado de las proteínas utilizando métodos computacionales.

Gracias a técnicas de cristalografía se ha logrado conocer las estructuras de varias proteínas Pero estas técnicas no se pueden utilizar para cualquier proteína; por eso se requiere una alternativa teórica para resolver su estructura.

Para obtener la estructura de las proteínas a partir de su secuencia se utilizaron distintas variantes de un método de optimización estocástico, llamado *recocido simulado*, que minimiza una función de energía. Como funciones de energía se utilizaron dos herramientas de alineamiento estructural; la más utilizada fue *Smith-Waterman*, que busca las similitudes en secuencia para aparear las dos cadenas. La otra, *SPalignNS*, alinea las estructuras de manera geométrica (no basándose en la secuencia). Luego de alinear las dos estructuras, se calcula la desviación media cuadrática (RMSD) para cuantificar su similitud.

En este trabajo se encontró que la optimización, utilizando *Smuth-Waterman*, tiene resultados buenos para estructuras cortas, menores a 200 aminoácidos, obteniendo un RMSD menor a 7 en un tiempo de optimización menor a 9 horas. Forzando los resultados a que tendieran al mínimo local más cercano, se redujo el RMSD hasta 3.17. En esta última etapa de refinamiento, las estructuras que contenían hélices alfa no se lograron mejorar tanto como las que no las tenían.

Además, se exploraron dos alternativas al método anterior utilizando; *SPalignNS* y una versión modificada del *recocido simulado* que copia valores del objetivo (substitución). *SPalignNS* no logró superar a *Smuth-Waterman* en ningún caso, aunque la diferencia no fue muy drástica: RMSD de 0.2 antes de refinar y a 1.3 después del refinamiento. El método de substitución mejora el tiempo de ejecución y RMSD para las cadenas largas.

Índice general

Capítulo 1

Introducción

Las proteínas juegan un papel clave en el funcionamiento de la célula. Estas se encargan de realizar prácticamente todas las funciones que permiten a la célula reproducirse, defenderse de agentes externos y alimentarse. Además son la unidad básica de muchas de las estructura que forman los organismos multicelulares. Por ello, entender las bases físicas del funcionamiento de las proteínas es un área de gran interés en la biología.

Hoy en día se sabe que todas las proteínas están compuestas por aminoácidos que se unen en una secuencia de longitud variada. La longitud (número de aminoácidos) de las proteínas puede ir desde un par hasta miles de aminoácidos. Éstas se forman en un proceso conocido como la síntesis protéica en la que una maquinaria (ribosomas) lee el orden y tipo de los aminoácidos, que están codificados en forma de tripletas de bases nitrogenadas (codones), del ácido ribonucleico mensajero (ARNm) y los va uniendo de uno en uno [1].

En la Naturaleza, los aminoácidos pueden ser de 20 tipos distintos, cada uno con diferentes tamaños y propiedades fisicoquímicas. Una vez que se sintetiza una proteína ésta se pliega adoptando una conformación estable dentro de su medio de acuerdo a las propiedades y las posiciones de los aminoácidos que la componen. Se postula que la estructura tridimensional que adopta la proteína está estrechamente relacionada con la función que ésta lleva a cabo dentro de la célula. Aunque la Naturaleza de la relación estructura-función de las proteínas no se ha catalogado formalmente [6], se reconoce que siendo la estructura y la función características del mismo objeto (proteína), deben estar relacionadas.

La estructura atómica tridimensional (3D) de las proteínas se infiere utilizando técnicas crista-

lográficas. Sin embargo, el método de cristalización de proteínas es una técnica en extremo complicada y no es posible obtener resultados satisfactorios con todo tipo de proteínas. Existen otros métodos además de la cristalografía para caracterizar la estructura de proteínas (i.e., microscopía electrónica, resonancia magnética nuclear, difracción de neutrones), pero a la fecha las técnicas de cristalografía han sido más exitosas para dicho fin [3].

Gracias a los avances en la secuenciación del ácido desoxiribonucleico (ADN) se conocen cientos de millones de secuencias de proteínas [21], pero sólo se han resuelto por el orden de las 100,000 estructuras de proteína [22].

Debido a esto, se han buscado métodos alternos para conocer el plegamiento final de una proteína dentro de un medio sólo a partir de su secuencia. A pesar de que se sabe que esto sucede espontáneamente a causa de las interacciones entre los distintos aminoácidos y el medio, no es posible resolver el problema de manera analítica, debido a la enorme importancia de las interacciones no locales el problema tiene una complejidad NP completo [1] [2, 3].

Para lograr atacar este problema se utiliza una amplia variedad de métodos de optimización computacional. Las técnicas y aproximaciones para resolver el plegado son muy variadas y se dividen en cuatro tipos [3]:

- Método de primer principio sin información de bases de datos: Basados en principios termodinámicos y que el plegado de las proteínas debe de ser un mínimo global, se asigna algún criterio de energía basado en las propiedades fisicoquímicas y se busca acercarse lo más posible al mínimo.

- Método de primer principio utilizando bases de datos: Apoyándose en la enorme base de datos de proteínas resueltas se buscan fragmentos recurrentes y se utilizan como piezas para formar la proteína desconocida. Las piezas se unen buscando minimizar un potencial (similar al método anterior).

- Reconocimiento de plegado: Varios estudios han identificado que el número de plegados finales de la mayoría de las proteínas es finito. Por esto se usan estructuras comunes como moldes a los cuales se busca ajustar las secuencias nuevas.

[1] Los problemas con complejidad NP (tiempo **P**olinomial **N**o determinístico) son aquellos problemas para los cuales no se conoce una solución que los resuelva en tiempo polinomail; pero dada una solución, es posible comprobar si es correcta en tiempo polinomial. Los problemas NP completos son una conjunto particular de NP, que si logran ser resueltos pueden ser usados para resolver todos los problemas de la clase NP.

- Alineamiento en secuencia y modelado comparativo: Asumiendo que una similitud en secuencia implica un plegado similar se hace la búsqueda utilizando alineamiento en secuencia con proteínas conocidas.

Actualmente se han popularizado versiones híbridas de estos métodos lo cual ha logrado superar las limitaciones que tiene cada método de forma individual.

En este trabajo se utilizan técnicas híbridas. Los resultados principales usan una estructura resuelta que se sospecha es similar, hacen un alineamiento en secuencia, basado en éste se crea una función de energía usando la distancia del modelo al molde y se utiliza un método de optimización para acercarse lo más posible al mínimo global.

Capítulo 2

Marco Teórico

2.1. Biológico

Las proteínas son la biomoécula más abundante y variada, con millones de tipos distintos y tienen una enorme gama de tamaños y formas. **Son el medio por el cual la información genética se manifiesta**, (i.e. su secuencia está codificada en la molécula de ADN por lo que forman parte de la herencia genética de los seres vivos). Se presentan en cualquier parte de la célula y participan en prácticamente todas las funciones de ésta [1].

La expresión génica es el mecanismo por el cual la información genética es expresada. En la figura 2.1 se muestra una diagrama de este proceso para una célula eucarionte.

Primero dentro del núcleo, el ARN polimerasa separa la doble hélice del ADN en la zona que se va a expresar y hace una copia complementaria de ARN. Es decir, se coloca la base nitrogenada complementaria de cada sito en el ADN y se substituye la timina por el uracilo.

Luego se "procesa" el ARN generado, removiendo zonas que no codifican para ningún aminoácido (intrones) y conservando las que sí (exones). A este proceso se le conoce como *splicing* (que se pude traducir al español como *acoplamiento*), al final del cual se obtiene lo que se conoce como ARN mensajero (ARNm).

El ARNm se exporta fuera del núcleo donde es transcrito a una proteína. En este proceso el ARNm se une a un ribosoma que, para cada 3 bases nitrogenadas, coloca el aminoácido correspondiente utilizando los ARN de transferencia (ARNt) del medio.

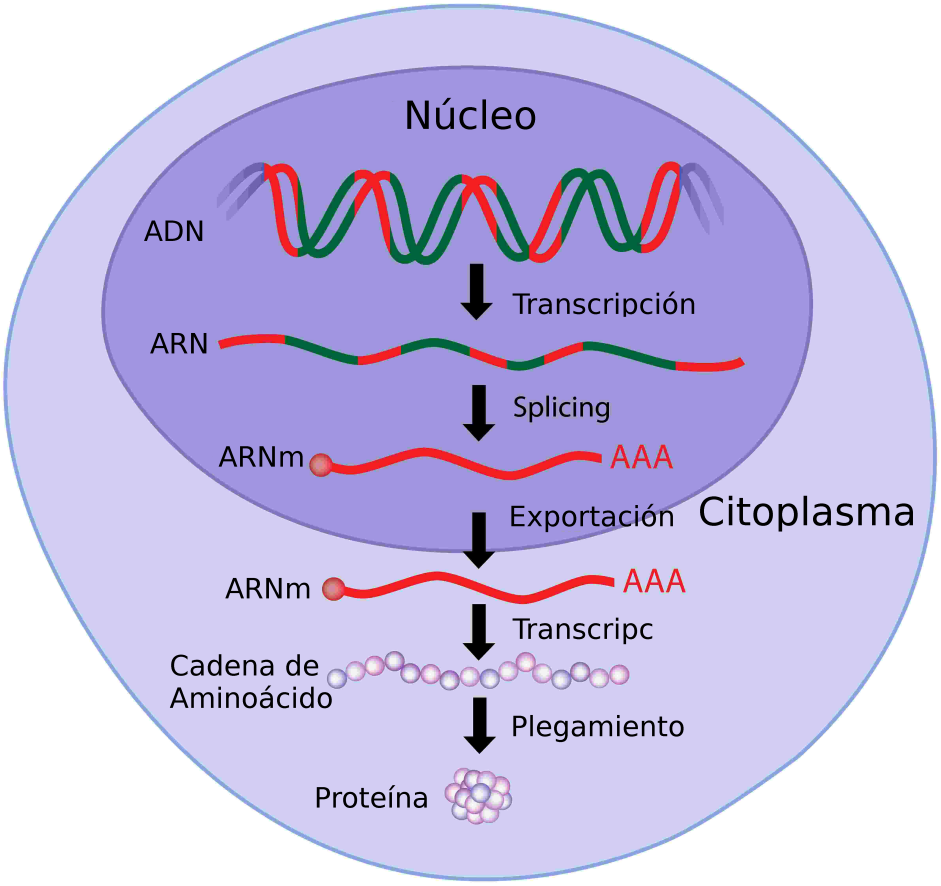


Figura 2.1. Diagrama donde se muestran los distintos pasos de la expresión génica para una célula eucarionte. [38]

2.1.1. Aminoácidos

Los aminoácidos se agrupan dentro de una misma categoría debido a la estructura común que comparten. Tienen un grupo carboxilo y un grupo amino, ambos unidos al mismo átomo de carbono central conocido como carbono alfa ($C\alpha$) [9]. En la figura 2.2 se puede ver la estructura de todos los aminoácidos.

Cuando se tienen cuatro compuestos distintos unidos a una misma molécula a ésta se le denomina un centro quiral. Aunque la conformación de la molécula es la misma, la configuración absoluta de estas puede ser de dos tipos: dextrógiras (D) y levógiras (L). Cada una es la imagen espejo de la otra y por eso son dos estructuras distintas [9]. En la figura 2.3 se puede ver la diferencia entre los dos tipos de configuraciones.

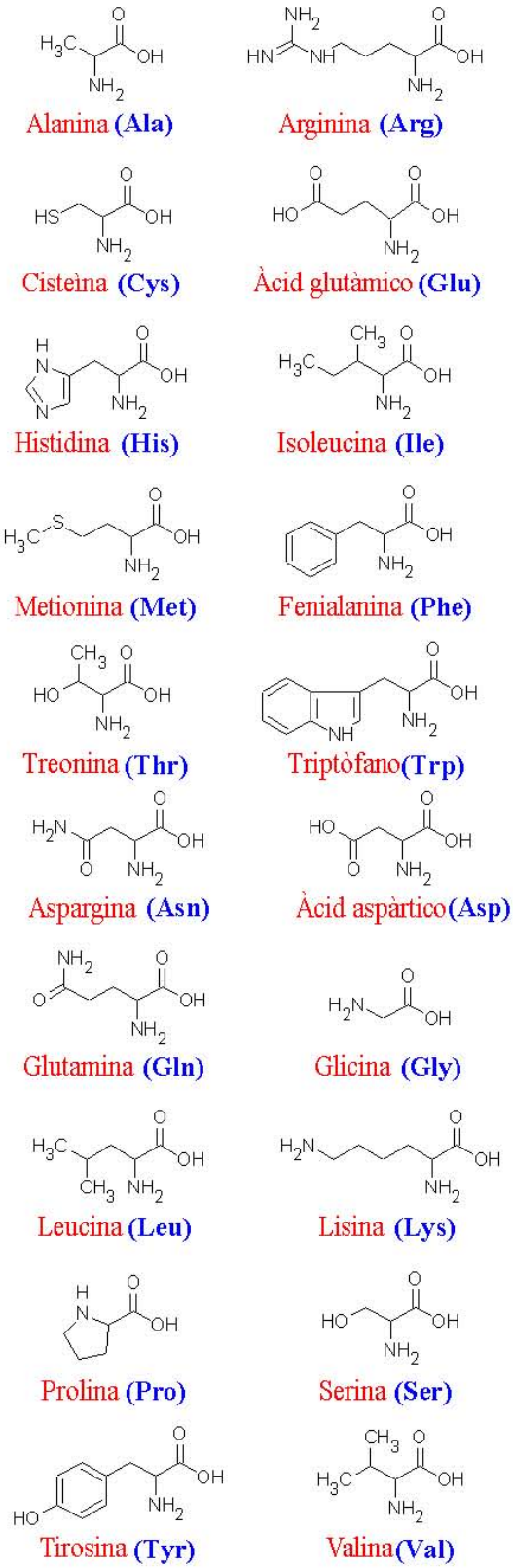


Figura 2.2. Estructura química de los veinte aminoácidos. [37].

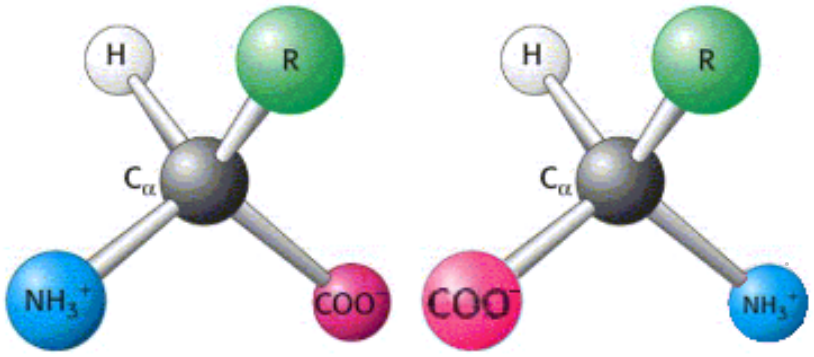


Figura 2.3. Configuración de un aminoácido tipo L (izquierda) y tipo D (derecha)

Cuando los aminoácidos son fabricados artificialmente se obtiene una mezcla de proporción igual de las dos quiralidades, y suele ser complicado distinguir entre las dos clases. Pero en la Naturaleza se presentan casi exclusivamente aminoácidos de tipo L. Algunos de los pocos ejemplos de aminoácidos con quiralidad D son péptidos presentes en el veneno del ornitorrinco y otros animales y las secreciones de algunos moluscos [4, 5].

Lo que distingue a cada uno de los veinte aminoácidos es su cadena lateral (indicada como R en las formulas químicas; como la figura 2.3 y 2.4) que sale del carbono central. Estas tienen distintos tamaños, formas, cargas y componentes que en conjunto afectan las propiedades fisicoquímicas de la molécula.

Para facilitar el trabajo con los aminoácidos, aprovechando su reducido número, se ha propuesto una nomenclatura para representarlos usando solo tres letras o sólo una. Esta se muestra en la tabla 2.1.

Las propiedades fisicoquímicas de los aminoácidos son de particular interés para la bioquímica, pues contribuyen en gran medida a la determinación del comportamiento y función de las proteínas. Es conveniente simplificar su descripción agrupándolos de acuerdo a estas propiedades:

No polares: alanina, valina, leucina, metionina, prolina, isoleucina y glicina (que carece de cadena lateral por lo que su efecto real es mínimo.

Aromáticas: fenilalanina, tirosina, triptofano. Relativamente no polares, forman puentes de hidrógeno, absorben luz ultravioleta.

Polares (hidrofílicos): serina, treonina, cisteína, aspargina, glutamina. Forman puentes de hidrógeno con agua

Positivas (básicas): lisina, histidina, arginina

Aminoácido	Abreviación	Símbolo
$Glicina$	Gly	G
$Alanina$	Ala	A
$Prolina$	Pro	P
$Valina$	Val	V
$Leucina$	Leu	L
$Isoleucina$	Ile	I
$Metionina$	Met	M
$Fenilalanina$	Phe	F
$Triosina$	Tyr	Y
$Triptófano$	Trp	W
$Serina$	Ser	S
$Treonina$	Thr	T
$Cisteína$	Cys	C
$Asparagina$	Asn	N
$Glutamina$	Gln	Q
$Lisina$	Lys	K
$Histidina$	His	H
$Arginina$	Arg	R
Ácidoaspártico	Asp	D
Ácidoglutámico	Cys	E

Tabla 2.1. Tabla con las representaciones de tres y una letra para los veinte aminoácidos.

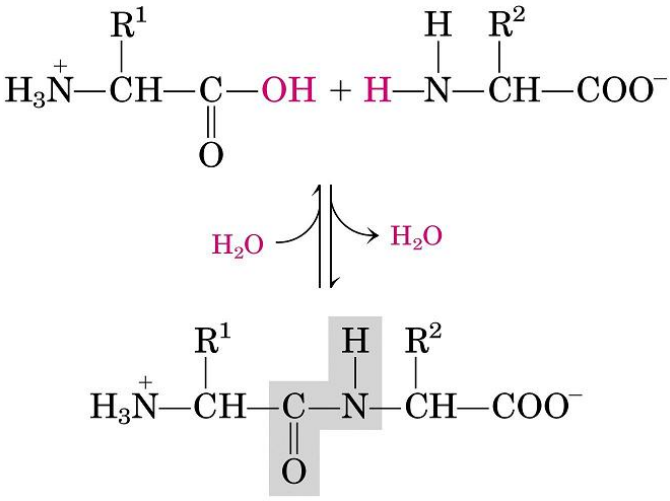


Figura 2.4. Diagrama ejemplificando la formación de un enlace peptídico [36].

Negativas (ácidos): ácido aspártico y ácido glutámico.

Esta es una de las muchas clasificaciones que hay para los aminoácidos. Existen otras que se basan en propiedades distintas como absorción de determinadas frecuencias de luz, peso molecular, presencia de elementos importantes como el azufre, etc [23].

2.1.2. Proteínas

Dos aminoácidos se pueden unir por un enlace covalente conocido como enlace peptídico; como se muestra en la figura 2.4. Éste se forma por deshidratación, remover una molécula de H_2O, tomando un grupo OH del extremo carboxilo y un H del extremo amino. A pesar de que la formación de enlaces peptídicos es una reacción exotérmica, ésta requiere una alta energía de activación, por esto los enlaces son sumamente estables con una vida media de aproximadamente 7 años [1]. La geometría y propiedades de este enlace se describen a mayor detalle en la sección 3.1.

Mediante este mismo proceso se pueden formar cadenas de aminoácidos de cualquier longitud. A las cadenas cortas se les denomina péptidos y a las cadenas más largas proteínas. No existe un consenso que defina claramente la frontera entre estas dos, un péptido suele tener aproximadamente una decena de aminoácidos mientras que una proteína tiene de las decenas en adelante [1].

A los fragmentos de aminoácidos que pierden ambos extremos al formar las proteínas se les conoce como residuos de aminoácidos o residuos. Al extremo que conserva su grupo amino se le llama amino-terminal y al que conserva su grupo carboxilo carboxilo-terminal.

Las proteínas pueden tomar formas muy complejas. Para facilitar su clasificación, descripción y estudio se han estandarizado distintos niveles de complejidad en las conformaciones nativas:

La estructura **primaria**, se refiere a los enlaces covalentes que unen la proteína, principalmente enlaces peptídicos. Esto incluye la secuencia de aminoácidos que compone la proteína.

La estructura **secundaria** esta compuesta por dos tipos de arreglos estables regulares los cuales se han observado que pueden formar las proteínas: las hélices alfa y las hojas beta. Se describen estas dos estructuras con mayor profundidad en la sección 2.1.3.

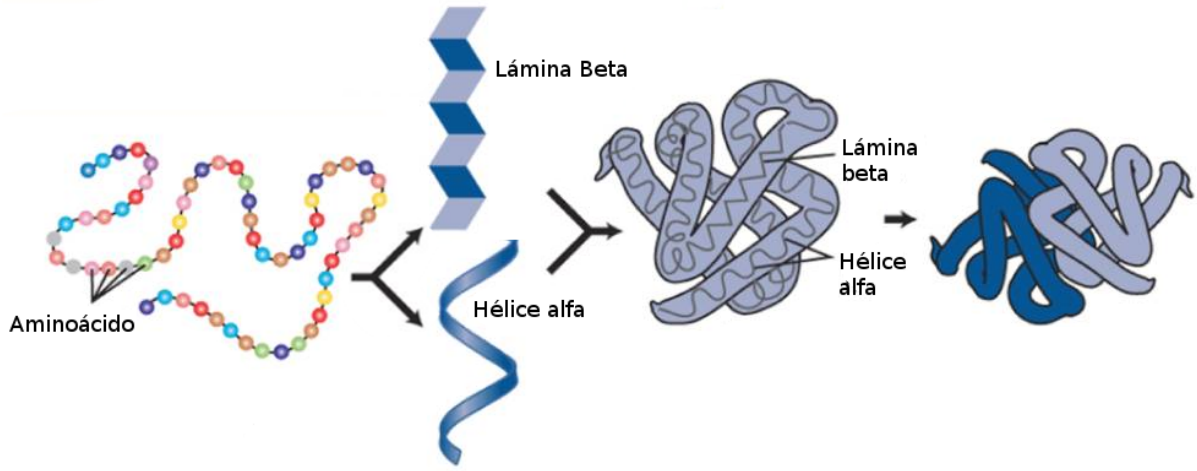


Figura 2.5. Los cuatro niveles de estructura de las proteínas [39].

La estructura **terciaria** es la forma tridimensional que toma la proteína en su plegado final.

Si la proteína está compuesta por dos o más cadenas, la forma como éstas se acomodan para obtener la proteína final se le conoce como estructura **cuaternaria**. No sólo es posible que varias proteínas se agrupen para formar un complejo más elaborado; también, una vez formadas, pueden ser adicionadas con grupos no aminoácidos, denominados grupos prostéticos, como lípidos, glucosa o metales para darles una función específica. A éstas se les conoce como proteínas conjugadas.

Cada proteína tiene una secuencia única de aminoácidos que determina su función.

Algunas de las observaciones que sustentan esta afirmación son:

- Se ha identificado que la causa de muchas enfermedades genéticas es provocada por mutaciones en la secuencia de aminoácidos de proteínas particulares (e.g., Huntington).
- Se ha observado que proteínas con funciones análogas en distintas especies suelen tener secuencias muy similares.

Aún así es importante notar que es posible variar la secuencia de una proteína sin afectar su plegado y función. Se sabe que aproximadamente un 30 % de las proteínas de los humanos tienen alguna variación y aún así conservan su función. Esto es porque la mayoría de las proteínas tienen zonas de poca importancia para su función; de la misma forma hay otras cuya presencia es esencial. No es fácil identificar estas zonas en las proteínas y, actualmente, hay muchos investigadores intentando resolver este problema [1].

2.1.3. Forma de las proteínas

Al conjunto de arreglos espaciales que puede tomar una proteína sin modificar ninguno de sus enlaces peptídicos se le conoce como **conformaciones**. En la sección 3.1.1 se describe qué movimientos son posibles dentro de la proteína y se da una idea matemática del espacio de conformaciones.

De entre todas las conformaciones posibles sólo unas cuantas se han observadas en condiciones experimentales (e.g., cristalográficas), a éstas se le conoce como conformaciones **nativas**. Se requiere que exista más de una conformación estable ya que se ha observado que ésta se modifica al unirse con otras moléculas o al reaccionar con elementos de la célula [1].

Es de esperarse que las conformaciones estables son las que minimizan la energía (energía libre de Gibbs), pero la diferencia entre una conformación nativa y una desdoblada es mínima (del orden de los 20 a 65 kJ/mol).

Las interacciones débiles que se dan entre distintos aminoácidos de la proteína (puentes de hidrógeno, interacción polar con el agua del medio) se pueden romper con poco gasto energético, de 4 a 30 kJ/mol. Para poner esto en perspectiva, los enlaces covalentes, como el enlace disulfuro, requieren más de 200 kJ/mol. Aún así los enlaces débiles son mucho más abundantes que los covalentes, su fuerza sumada termina opacando la de los enlaces fuertes y se convierte en un factor importante dentro de la proteína.

Pero, dentro de una proteína, durante el proceso de plegamiento de su estructura terciaria o cuaternaria, las interacciones de los aminoácidos con el agua se sustituyen por interacciones con otros aminoácidos, por lo que el cambio neto de energía resulta ser muy cercano a cero, así su contribución para estabilizar las conformaciones nativas no es grande [1].

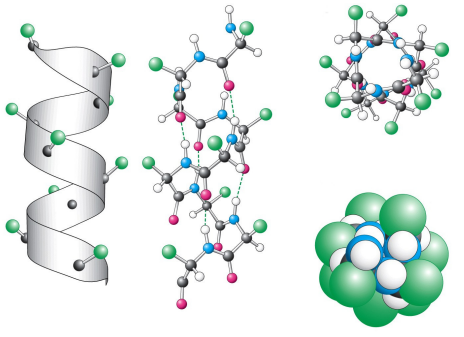
Como se puede ver, el determinar una fuerza primordial que provoca la estabilidad de una con-

formación nativa no es sencillo. Para analizar este problema de forma alternativa se pueden estudiar las estructuras secundarias.

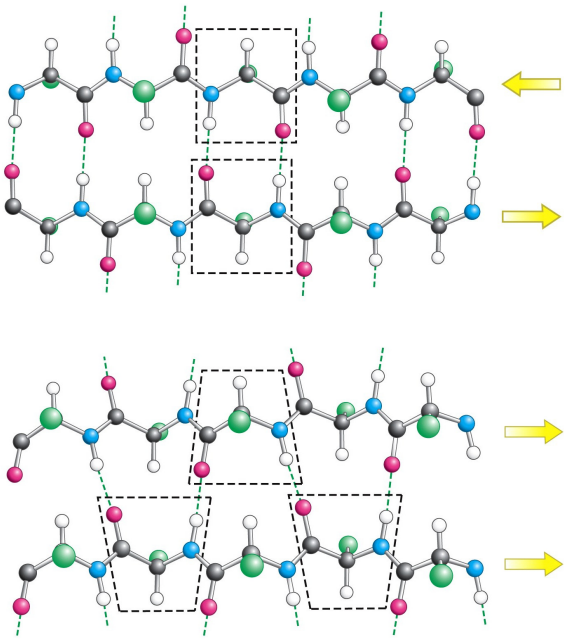
Las estructuras secundarias son patrones recurrentes que se encuentran en la mayoría de las proteínas. Se ha observado que maximizan el número de interacciones débiles

En la hélice alfa (figura 2.6a) la cadena polipeptídica se enrolla alrededor de un eje imaginario dejando los grupos R de los residuos hacia afuera (alejándose del eje). La unidad mínima de esta estructura es una vuelta entera de la hélice con una altura aproximada de 5.4 Å y abarca 3.6 aminoácidos. Los ángulos diedros que determinan esta estructura son en promedio $\psi = -45$ y $\phi = -60$. En promedio, una cuarta parte de todos los aminoácidos de las proteínas forman parte de una hélice alfa, porque es la estructura que mejor optimiza los enlaces débiles a nivel local [1].

Las hojas beta (figura 2.6b) son estructuras en forma de zig-zag que se alargan sobre un plano, varias de éstas pueden ser colocadas en paralelo formando una estructura plana. Los grupos R de los residuos son normales al plano que forman y se alternan por arriba y por debajo. Dependiendo de cómo se acomoden las cadenas adyacentes pueden ser paralelas o antiparalelas, cada una tiene un periodo distinto: 6.5 Å y 7 Å respectivamente [1].



(a) Hélice alfa.
A la izquierda vista latera, a la derecha vista superior.



(b) Hoja beta.
Arriba acomodo paralelo, abajo acomodo antiparalelo.

Figura 2.6. Diagrama de las configuraciones de las estructuras secundarias de las proteínas [34].

Actualmente se cree que para las proteínas globulares, aquellas que se pliegan en medio acuoso, la principal fuerza que estabiliza las conformaciones nativas es el llamado **colapso hidrofóbico**.

Si una proteína globular está completamente desdoblada tiene su superficie expuesta al medio acuoso. Los aminoácidos hidrofóbicos tienden a ser rodeados por las moléculas de agua; este acomodo alrededor de la proteína disminuye la entropía del agua. Cuando la proteína se pliega de una forma más compacta ciertas zonas son completa o parcialmente cubiertas, reduciendo la cantidad de la molécula que está expuesta al medio. Se puede ver que si la cantidad de hidrófobos expuestos disminuye, las moléculas de agua rodeándolos será menor y se aumentará la entropía del medio.

2.2. Cristalografía de Proteínas

Desde 1934, J.D. Bernal y Dorothy Hodgkin [16] observaron que los cristales de proteínas difractaban rayos X produciendose un complicado patrón de difracción. Esta información no pudo ser descifrada hasta 1958 cuando J. C. Kendrew et al. [17] aplicaron la técnica desarrollada por Max Ferdinand Perutz [18], que hacía posible unir átomos pesados a distintas zonas de una proteína. La enorme diferencia de carga de estos elementos pesados comparada con el resto de la proteína provocaba un patrón de difracción suficientemente distinto del normal el cual se podía utilizar para hacer un análisis estructural de la molécula. Con esto se logró obtener la densidad electrónica de una mioglobina a partir de la cual se produjo una imagen tridimensional de ésta. El método que Kendrew utilizó sigue siendo usado hoy en día para conocer la forma de las conformaciones nativas, es un proceso bastante complicado y con muchas restricciones [25].

2.2.1. Descripción general

Para conocer la conformación nativa de una proteína es necesario un largo y complicado proceso. Primero se purifica la proteína a estudiar, luego se cristaliza utilizando varios métodos químicos. La cristalización de proteínas es un proceso casi artesanal. Para proteínas de las cuales se desconoce por completo su estructura, en contraste con algunas para las cuales se sabe o sospecha que tendrán un plegado similar al de una proteína conocida, es necesario hacer varios cristales cada uno con distintos dopajes de átomos pesados.

Una vez fabricados los cristales se les ilumina con rayos X. Debido a las dimensiones atómicas

de las proteínas es necesario utilizar luz de frecuencia equivalente a la separación de sus componentes ($\lambda \approx 1.5418$Å en el rango de los rayos X) para lograr verlas con claridad. Esto produce un complicado patrón de difracción que tiene en promedio unos 5000 puntos, esto se repite para aproximadamente 300 ángulos de incidencia distintos. Si son necesarios se repite el proceso para cada uno de los cristales dopados.

Finalmente toda esta información se procesa por computadora para obtener una imagen de densidad electrónica de la proteína. Ésta simplemente indica qué forma tiene la proteína, es necesario ajustar la posición de cada átomo y el enlace entre estos a la densidad electrónica para conocer el plegado que ésta tiene [25, 33].

2.2.2. Rayos X

Los rayos X interactúan con las partículas cargadas (protones y electrones) de la proteína, éstas vibran con la misma frecuencia de la onda incidente por lo que emiten una onda de la misma frecuencia con una probabilidad proporcional a la amplitud de la onda incidente. Debido a que los electrones tienen una carga mucho mayor relativa a su masa en comparación a los protones, casi la totalidad de la onda reflejada (emitida) proviene de los electrones.

Las ondas que emiten los átomos de una sola proteína son de muy baja intensidad y se pierde en el ruido de fondo. Para resolver esto se utilizan cristales donde la repetición periódica del acomodo cristalino amplifica la señal. Otra interpretación equivalente es que en un cristal se tiene toda una familia de planos de Bragg (ver apéndice B para más detalles sobre estos planos), todos paralelos y separados por la misma distancia. Juntos provocan la misma diferencia de fase de los rayos incidentes por lo que se amplifica la intensidad de salida [33]. Por esto tiene sentido describir la distribución de electrones con funciones periódicas.

A diferencia de cuando los rayos X inciden sobre un cristal simple donde la onda se refleja en unos pocos planos de Bragg, como la proteína no tiene una estructura trivial, un sólo rayo incidente se refleja en varios cientos de planos de Bragg distintos, el patrón de interferencia resultante es extremadamente complicado, como se puede ver en la figura 2.7.

Cada punto brillante corresponde a un distinto plano de Bragg del cristal [27]. Una explicación más detallada de la relación entre el cristal de proteína y el patrón que forma se encuentra en el apéndice C.

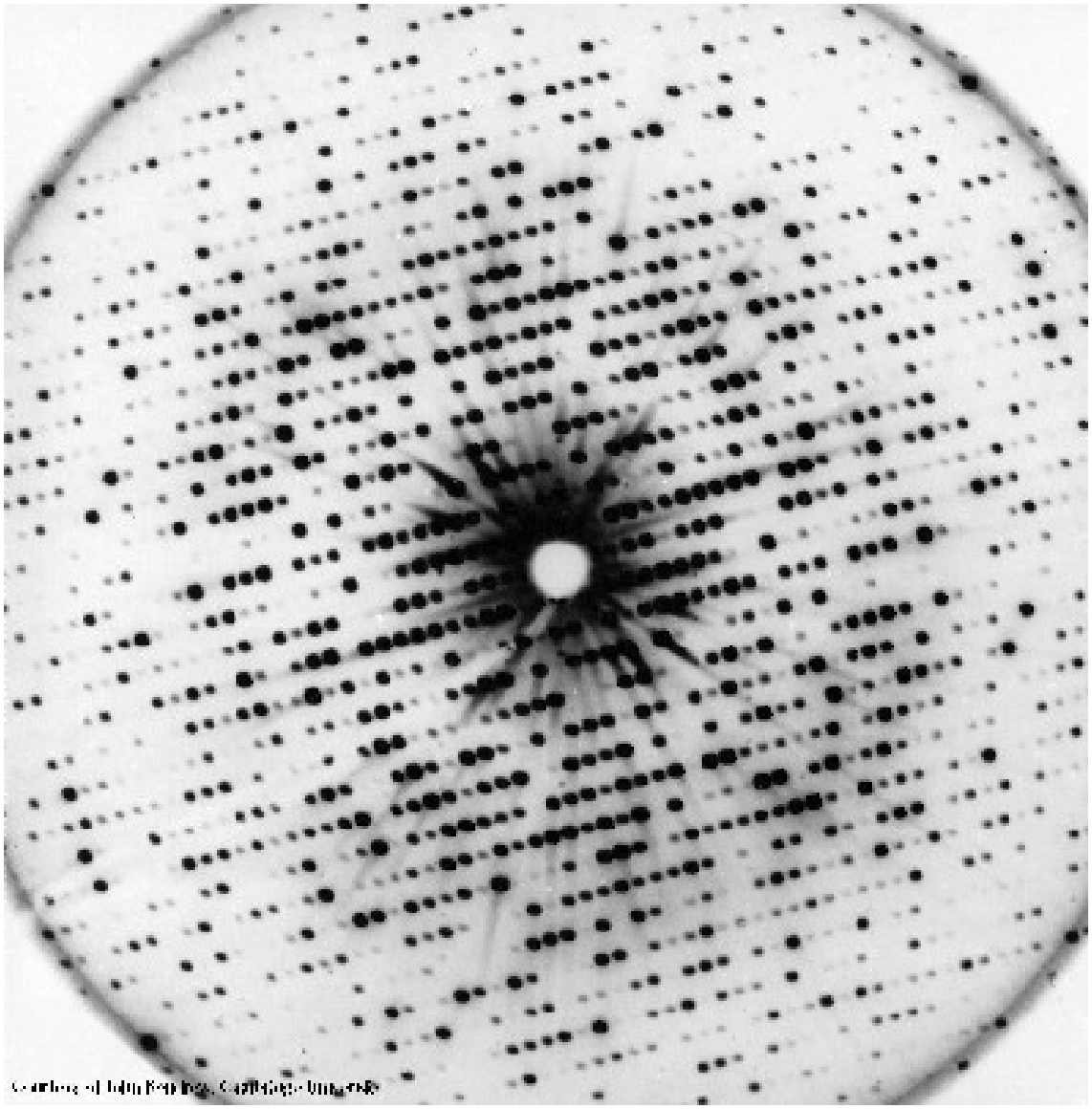


Figura 2.7. Patrón de difracción producido por un cristal myoglobina de ballena, mismo que fue usado para resolver la primera proteína por Kendrew et al. [35]

Este patrón contiene mucha información sobre la distancia relativa y ubicación de los átomos. Las ondas emitidas por cada átomo terminarán interfiriendo constructiva o destructivamente dependiendo de la distancia entre ellos. Si sólo se tratase de unos cuantos átomos se podría fácilmente hacer una regresión para conocer las distancias relativas de las fuentes emisoras, pero en el caso de las proteínas es necesario hacer múltiples mediciones y cálculos sumamente elaborados.

2.2.3. Interpretación del patrón

Sea $\rho(x, y, z)$ la función que describe la distribución de densidad electrónica de un cristal. Se supone por simplicidad que la estructura se repite cada a unidades en dirección x, b unidades en dirección y y c unidades en z. Esto corresponde a un cristal con una celda unitaria paralelepípeda con lados $a \times b \times c$.

Así, la función ρ será triplemente periódica (con periodos a, b, c) y consecuentemente puede ser representada como una triple serie de Fourier:

$$\rho(x, y, z) = \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{hkl} e^{-i(\frac{2\pi h}{a}x + \frac{2\pi k}{b}y + \frac{2\pi l}{c}z)}. \tag{2.1}$$

en la ecuación 2.1 los índices hkl coinciden con los índices racionales del cristal[1], así cada triplete determina un conjunto de planos de Bragg los cuales provocan la reflexión C_{hkl}.

A C_{hkl} se conoce como el factor de estructura. Este valor es muy importante en la cristalografía ya que, como se puede ver en la ecuación, si se conoce el valor complejo C_{hkl} entonces es posible calcular ρ y así resolver la estructura de la proteína.

Pero no es posible observar la contribución de todos los planos. Dependiendo de la orientación del rayo incidente con respecto al cristal algunos planos reflejarán la onda; esto, es el patrón de salida que se observa. Si el cristal se rota, el rayo será reflejado por otro conjunto de planos y el patrón será diferente.

Como ya se mencionó antes, cada punto del patrón de difracción corresponde a la reflexión sobre una familia de planos de Bragg y la intensidad del punto corresponde a $|C_{hkl}|^2$. Esto solo da una parte de la información requerida pera calcular $\rho(x, y, z)$, por desgracia no es posible conocer

[1] Los índices racionales del cristal ($m = (m_1, m_2, m_3)$) definen una familia de planos paralelos dentro del cristal. Se relacionan con los índices de Miller ($h = (h_1, h_2, h_3)$) de la siguiente manera: $h_i = \frac{t}{m_i}$. donde t es una constante arbitraria; normalmente se elige como el máximo común denominador de m_i [10]

el valor de la fase directamente del patrón de difracción. A esto se le conoce como el problema de la fase y es lo que J. C. Kendrew et al. lograron resolver utilizando un técnica denominada *Multiple Isomorphous Replacement* (MIR) [26].

2.2.4. Multiple Isomorphous Replacement

Esta es la técnica originalmente utilizada para resolver el problema de las fases. [25]. Se utiliza cuando se desconoce por completo la forma que tiene la proteína por resolver. Requiere tres cristales distintos, uno con la proteína pura y otros dos con la proteína dopada con átomos pesados en distintos sitios.

Los átomos pesados se colocan en el mismo sitio para todo un conjunto de proteínas, así, al cristalizar, las propiedades periódicas no se pierden. Además, usualmente sólo se tiene un tipo átomo pesado por proteína y este se coloca siempre en los mismos sitos se puede conocer fácilmente el tipo de estructura cristalina que todos forman (simple cubic, face center cubic, etc). Con esta información se puede calcular la amplitud y fase de todas las reflexiones que este produce.

El átomo pesado, al tener muchos más electrones que el resto de la proteína, provoca una dispersión mucho más intensa y modifica notoriamente la fase e intensidad de todo el patrón de difracción. Con esta información es posible calcular la fase del cristal original.

En la figura 2.8 F_P es la onda generada por el cristal de proteína pura, F_{PH} la generada por un dopaje de átomo pesado, ambas tienen un círculo de la misma longitud representando la incertidumbre de su fase. F_H representa la onda que se conoce del cristal teórico de sólo átomos pesados. Utilizando ésta información se puede trasladar el círculo F_{PH} con el vector F_H y reducir el problema a sólo dos fases, pero si se hace un segundo dopaje (ver figura 2.9), F_{H2}, es posible repetir el mismo proceso y así determinar por completo la fase [25, 28, 33].

2.3. Cómputo

Para utilizar un método de optimización, es necesario asignarle una *energía* o puntuación a cualquier modelo fabricado para representar qué tan bueno es, ya que este valor es el que el método busca optimizar. Por esto la función de energía que se utiliza es determinante para el resultado; la función determina la forma del espacio de búsqueda. Una función que resulta ser muy suave puede

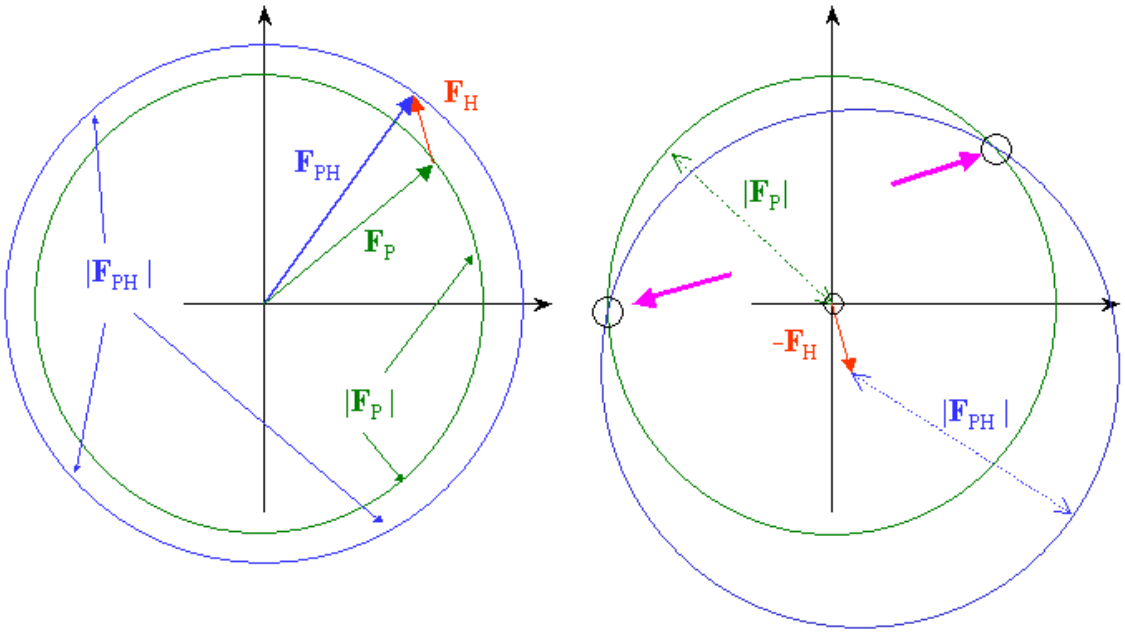


Figura 2.8. Si sólo se conoce la amplitud de la ondea (radio) en la representación polar se tiene un círculo. A la izquierda se muestran $|F_P|$, $|F_PH|$ y F_H. A la derecha se utiliza F_H para trasladar uno de los círculos y reducir el problema de la fase a dos puntos. [28]

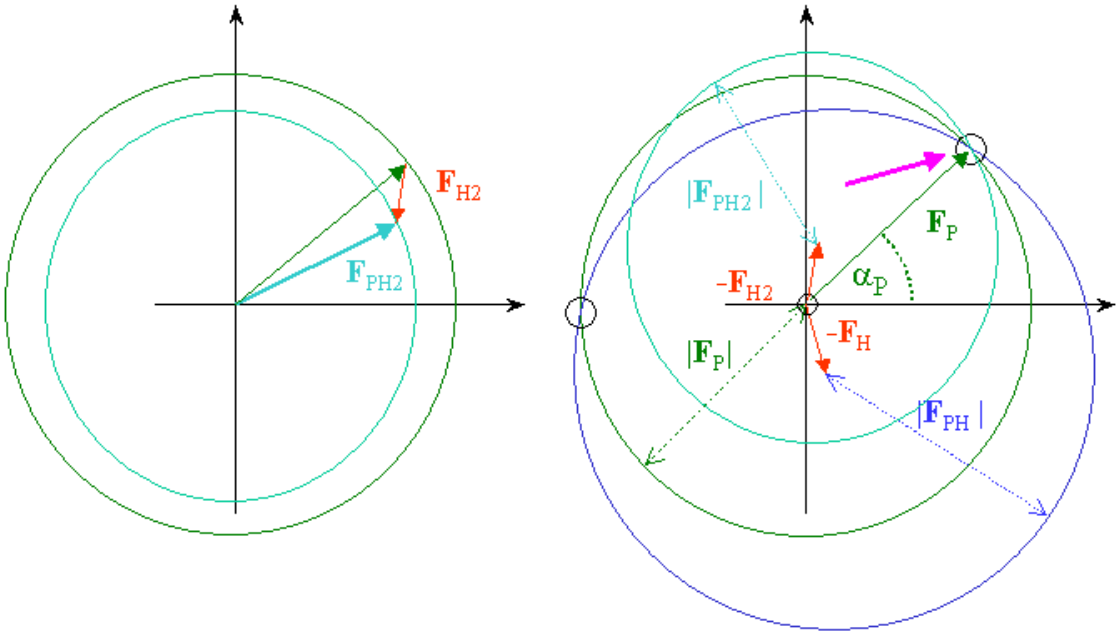


Figura 2.9. Con un segundo dopaje se tienen tres círculos y dos vectores para desplazarlos, por lo que, como se ve en la imagen de la derecha, se puede obtener un único punto. [28]

facilitar la búsqueda así como una muy rugosa puede entorpecerla.

Puede parecer evidente, pero la función de energía debe reflejar claramente el parecido del modelo con la realidad. Si la función de energía no logra reflejar un cambio ligero en el modelo como un cambio ligero de energía o si la diferencia de energía entre dos objetos [2] que se sabe que son similares es mayor que la de otros dos que son claramente distintos, la función no es apta para una optimización. Si esto se elige mal, no sólo puede entorpecer, sino imposibilitar por completo la búsqueda con cualquier método de optimización.

Es necesario tener una manera de representar el objeto a optimizar numéricamente, para así introducirlo en la función de energía. Al hacer esto se debe buscar tener la menor cantidad de entradas, siempre y cuando sean suficientes para representar todos los atributos importantes del objeto a optimizar.

A la sobrerepresentación de un atributo se le conoce como redundancia. La redundancia puede afectar la optimización de dos formas: si un sólo atributo está sobrerepresentado su valor puede opacar al resto, o bien, si la sobrerepresentación no es tan abrumadora simplemente la búsqueda no será tan eficiente porque habrá variables innecesarias introduciendo ruido. Además, el tamaño del espacio de búsqueda afecta directamente la velocidad del método; un espacio pequeño implica una búsqueda más rápida y viceversa.

El pre-procesamiento que se le hace al objeto para extraer o transformarlo en los atributos de entrada debe elegirse con cuidado para que estos reflejen de manera útil los cambios en los factores que se van a optimizar, esto determina muy fuertemente la eficiencia de la búsqueda.

2.3.1. Gradiente Descendiente

El gradiente descendiente o gradiente en descenso es uno de los métodos de optimización más simples e intuitivos.

Este puede resolver una función diferenciable desconocida. En cada paso, se obtiene la derivada del punto actual y se avanza en dirección del vector derivada. Eso se repite hasta que eventualmente la búsqueda se estaciona en un mínimo. Este método trabaja muy bien para funciones que tiene un gran mínimo global, ya que el vector derivada siempre guiará hacia ese valor, pero resulta muy

[2] Durante toda la sección de cómputo se describe el problema de optimización de forma general refiriéndose como *objeto* a lo que se optimiza. En el caso de este trabajo el objeto a optimizar siempre serán los modelos de proteínas.

poco efectivo para funciones rugosas con múltiples mínimos locales. En estos, casos el gradiente descendiente se estanca en el mínimo más cercano sin posibilidad de encontrar el mínimo global. Para resolver esto, es común repetir la búsqueda varas veces; cada vez se comienza en distintas posiciones y conservar el mínimo valor obtenido

2.3.2. Recocido Simulado

Otro método de optimización muy sencillo pero a la vez muy eficiente es *simulated annealing* o *recocido simulado*. Éste, es una modificación del algoritmo *Metropolis* [14], se inspira en el método de metalurgia que lleva el mismo nombre: *recocido* (annealing), en la que un metal se calienta hasta su punto de fusión y luego se enfría muy lentamente favoreciendo así la formación de estructuras cristalinas [12, 13].

El método computacional utiliza una temperatura ficticia que le asigna al sistema, luego lo enfría lentamente hasta llegar a un estado de equilibrio (normalmente temperatura 1). Se ha demostrado que este método siempre converge al mínimo global si se enfría suficientemente lento y se explora el espacio con mucha aleatoriedad [12].

En cada paso, el método evalúa varios puntos al azar dentro de un área alrededor de su estado actual; si encuentra uno donde la energía sea menor que la de su estado actual, toma ese nuevo punto como su mejor candidato y comienza a buscar alrededor de éste. Lo que hace eficiente este método, y donde juega un papel la temperatura, es que tiene la posibilidad de aceptar un punto que aumente la energía del sistema de acuerdo a la ecuación:

$$P = e^{\frac{E_0 - E_1}{T}} = e^{\frac{-\Delta E}{T}}. \tag{2.2}$$

donde P es la probabilidad de aceptar la nueva solución, E_0 es la energía actual, E_1 la energía del nuevo punto y T la temperatura actual. Esta probabilidad esta basada en la distribución de Boltzmann, la constante k es eliminada porque, en este caso, la energía y la temperatura no tienen unidades y se pueden dividir directamente.

Es importante notar que esta fórmula sólo se utiliza cuando $E_0 < E_1$ por lo que $\Delta E < 0$. De esta forma, al comienzo de la búsqueda, cuando la temperatura es más alta, el cociente $\Delta E/T$ es cercano a 0 por lo que es bastante probable aceptar soluciones que no mejoren, pero a medida que

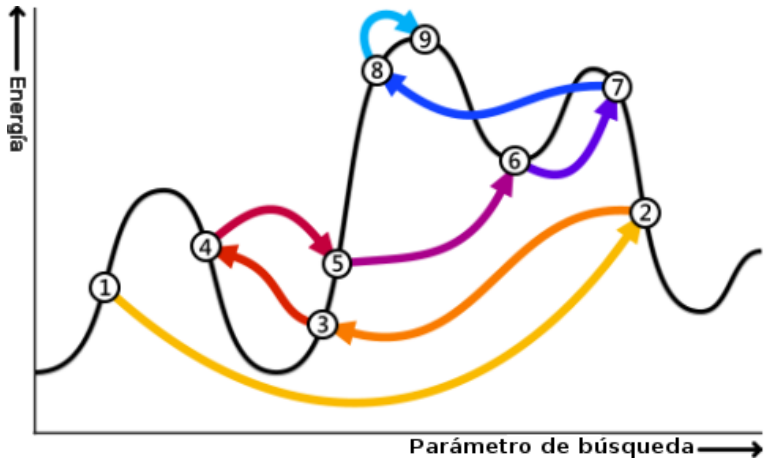


Figura 2.10. Diagrama que ejemplifica los pasos de una búsqueda sobre una función de una dimensión utilizando recocido simulado. Al inicio el método se mueve grandes distancias (1 a 3) y conforme avanza el tiempo el intervalo que explora se reduce [13].

la búsqueda progresa es cada vez más exigente.

Así, al inicio, el método permite explorar con bastante libertad el espacio en las cercanías *escalando* pendientes de energía y así, potencialmente, escapando de mínimos locales. Una vez que la temperatura disminuye, la búsqueda comienza a converger a un mínimo, rechazando las soluciones que no mejoran la energía actual.

Se espera que cuando la temperatura aún es alta, la búsqueda pase por encima de todos los mínimos locales y se estanque en un mínimo global, ya que estos son más grandes y profundos.

Por esto, cuando el espacio es rugoso (abundante de mínimos locales) este método es muy útil y supera a los métodos de gradiente que terminarían estancándose en el mínimo local más cercano de su punto de inicio.

Una vez elegida la función de energía, como en todos los problemas de optimización, hay algunos factores que se deben ir ajustando de manera que la búsqueda sea más eficiente.

2.3.3. Factores a optimizar

Como recurso pedagógico, se suele introducir a los estudiantes a muchos métodos de optimización usando problemas sencillos con soluciones conocidas. Esto hace muy fácil el ajuste de algunos factores libres. Si se conoce como trabaja el método que se utiliza y la forma del espacio de búsqueda, resulta muy sencillo ajustar intuitivamente estos factores. Pero en problemas reales la forma del espacio se desconoce casi por completo, de lo contrario el método de optimización sería inne-

cesario. El ajuste de los parámetros se hace por prueba y error o, si se cuenta con experiencia, de manera empírica.

Si se observa el cociente de la exponencial en la ecuación 2.2 es claro que la temperatura y la energía deben de ser del mismo orden de magnitud. Si la temperatura es muy alta, cualquier solución se aceptaría sin importar de la diferencia de energías, o en caso contrario todas serían rechazadas.

El número de ciclos de búsqueda, temperatura inicial y número de subúsquedas antes de disminuir la temperatura son las tres variables más importantes que se ajustan en *recocido simulado*.

Como ya se vio antes, la temperatura determina la probabilidad de aceptación de soluciones malas, las cuales, junto con el factor de enfriamiento, son las que indirectamente determinan el número de ciclos que durará la búsqueda.

La temperatura inicial debe de ser elegida, de tal manera que dada una diferencia de energía ΔE, que se considere aceptable para el objeto a optimizar, la probabilidad de aceptar esta solución sea razonable. Aquí se puede ver claramente un problema con éste y muchos métodos de evaluación numéricos, ya que no existe ningún criterio único o preestablecido para determinar qué diferencia de energía y probabilidad sean *aceptables*. Esto se debe asignar de manera intuitiva basado en el conocimiento que se tenga de la función de energía o simplemente por medio de prueba y error.

El número de ciclos sólo afecta la eficiencia de la búsqueda si es muy pequeño, aun así es importante notar que cuando la temperatura alcanza un cierto valor bajo, las soluciones aceptadas solo van a tender hacia el mínimo más cercano y, una vez que se alcance este mínimo, todos los ciclos de búsqueda siguientes son inútiles.

Si lo que se busca es un método rápido, se debe saber a partir de qué cantidad de ciclos la búsqueda deja de mejorar y se estanca. Una vez más, es muy difícil determinar este valor a priori, y en casi todos los casos es necesario experimentar con varias cantidades para así ir determinando el valor óptimo.

Cuando el espacio de búsqueda es muy grande y complicado el, punto inicial se convierte en un factor importante. Si éste se encuentra muy lejos del mínimo global o si el *camino* hacia éste mínimo es errático y rugoso, la probabilidad de que se llegue a encontrar la mejor solución se vuelve ínfima. Una solución, un poco similar a la fuerza bruta pero eficiente en estos casos, es hacer muchas búsquedas con puntos iniciales diferentes, así se logra explorar una mayor cantidad del espacio y se aumenta la probabilidad de acercarse al mínimo global. Si es importante la velocidad

de la búsqueda se puede reducir el número de ciclos de cada búsqueda para que el total de ciclos no sea tan grande, esto no afecta en forma notable el resultado, ya que lo que se pierde en precisión de cada búsqueda en particular se espera recuperar abarcando una mayor cantidad del espacio.

Capítulo 3

Método

3.1. Representación numérica del plegado de una proteína

Aunque es posible representar las proteínas como se presentan en los archivos de la *Protein Data Bank* (PDB) [22], es decir con las coordenadas en $\mathbb{R}^3$ de todos los átomos que las componen, sería una manera muy redundante y poco descriptiva que sólo entorpecería, sino es que haría imposible una búsqueda. Es necesario aprovechar las características conocidas de las proteínas para simplificar la representación y así reducir el espacio de búsqueda.

La estructura que comparten todos los aminoácidos, es decir el amino (NH_2) y el ácido carboxílico ($COOH$) unidos entre sí con un átomo de carbono, al igual que muchas moléculas, tiene una geometría bien estudiada.

Por esto, la posición de muchos de los átomos que la componen resulta redundante ya que se puede deducir con sólo conocer la de uno solo de sus componentes y la orientación de la molécula. En contraste, las cadenas laterales que distinguen a los diferentes aminoácidos no tienen una estruc-

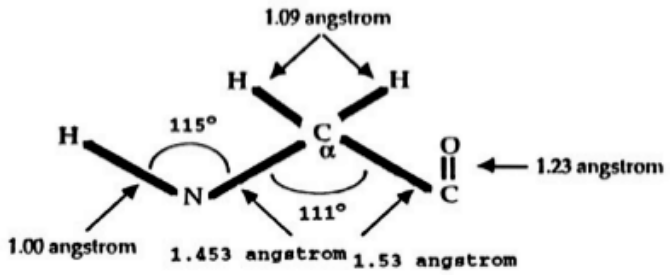


Figura 3.1. Geometría de un aminoácido. [7]

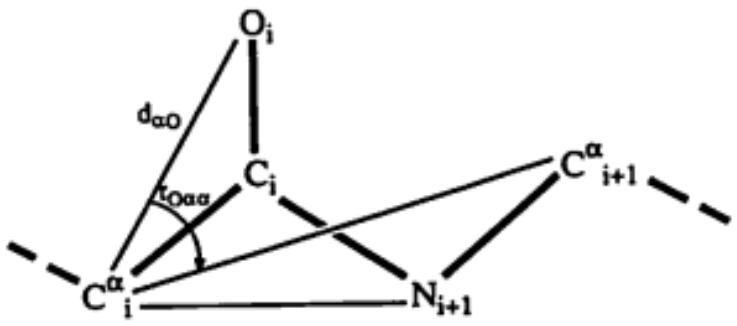


Figura 3.2. Posición del oxígeno de acuerdo a la orientación del aminoácido al que pertenece y el siguiente [8].

tura única; se le conoce como *rotámeros* al conjunto de formas que puede tomar la cadena lateral de un aminoácido.

Al estudiar las vastas muestras cristalográficas de proteínas se ha encontrado que en la estructura nativa de las proteínas los rotámeros sólo presentan un número finito de formas distintas. La cantidad de rotámeros que tiene cada aminoácido depende del tamaño y complejidad de la cadena lateral; por ejemplo, la alanina, cuya cadena lateral es CH_3, sólo tiene un rotámero, mientras que la lisina, con una cadena lateral mucho más compleja $(C_4H_{11}N)$, cuenta con más de 80 rotámeros conocidos. [24]

Finalmente, la posición del oxígeno de doble enlace (que no se pierde al hidrolizar los aminoácidos) está determinada, no por la orientación del mismo aminoácido al que pertenece, sino por el ángulo que forma con los átomos del siguiente aminoácido unido por el extremo del carboxilo.

La posición que toma el oxígeno del carboxilo con respecto a la orientación del siguiente aminoácido está mejor determinada por la distancia del carbono alfa al oxígeno y el ángulo $O_iC_{\alpha i}C_{\alpha i+1}$. Estas constantes son diferentes para cada aminoácido. [8]

Debido a la gran similitud entre todas las constantes conservadas para cada aminoácido (que se pueden ver en la tabla 3.1) se utilizó un promedio de cada constante para fijar la posición del oxígeno en todos los aminoácidos. Con todo esto se logra reducir la cantidad de variables necesarias para representar cada aminoácido.

También es posible hacer una reducción para toda la proteína; al igual que con los aminoácidos, los enlaces peptídicos que se forman entre ellos no pueden tomar cualquier posición. Para describir estas restricciones es necesario entender que son lo ángulos diedros.

Residuos	d_{CN}/Å	$\tau_{\alpha CN}$	$\tau_{CN\alpha}$
Gly	1.322	115.7°	121.2°
Ala	1.322	116.2°	121.4°
Ser	1.321	115.8°	121.3°
Cys	1.322	115.9°	121.3°
Val	1.321	115.8°	121.5°
Thr	1.322	116.0°	121.2°
Ile	1.320	115.6°	121.4°
Pro trans	1.321	115.8°	121.8°
Pro cis	1.323	116.5°	124.6°
Met	1.322	115.8°	121.4°
Asp	1.322	115.9°	121.6°
Asn	1.323	116.0°	121.5°
Leu	1.319	116.1°	121.5°
Lys	1.321	115.9°	121.7°
Glu	1.322	116.0°	121.4°
Gln	1.322	116.0°	121.6°
Arg	1.322	116.0°	121.2°
His	1.322	116.0°	121.3°
Phe	1.322	115.6°	121.5°
Tyr	1.323	115.6°	121.5°
Trp	1.326	116.0°	121.5°
Cyx	1.319	115.6°	120.9°

Tabla 3.1. Valor que determinan la posición del oxígeno para cada tipo de aminoácido [8].

3.1.1. Ángulos diedros

A diferencia de lo que sucede en $\mathbb{R}^2$, en $\mathbb{R}^3$ tres puntos no determinan un sólo ángulo; es posible determinarlo con cuatro puntos de las siguiente forma: En $\mathbb{R}^3$ cualesquiera tres puntos determinan un plano en el espacio. Así con cuatro puntos $p1, p2, p3, p4$ se tienen dos planos, el formado por $p1 - p2 - p3$ y el formado por $p2 - p3 - p4$. Estos planos, salvo si son paralelos, siempre van a intersectarse en una recta la cual cruza por los puntos $p2$ y $p3$. Al ángulo que forman los dos planos se le conoce como ángulo diedro.

Una vez que se forma el enlace peptídico, y se elimina el grupo amino y el ácido carboxílico, la parte del aminoácido que se conserva, ignorando los hidrógenos, es la cadena $N - C - C_\alpha$, la cual se le conoce como el esqueleto o *backbone* de una proteína.

La unión de dos aminoácidos determina tres ángulos diedros:

- $\psi \to N - C_{\alpha 1} - C - N_2$
- $\phi \to C_{-1} - N - C_\alpha - C$

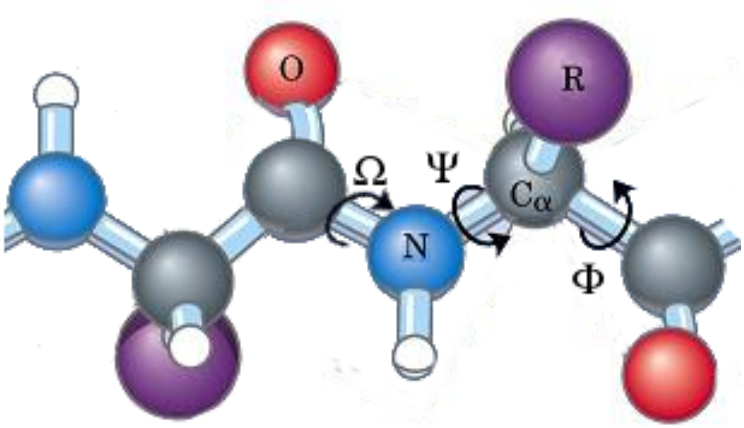


Figura 3.3. Los tres ejes de rotación permitidos en un enlace peptídico; definidos por los rotámeros ψ, ϕ, ω.

- $\omega \rightarrow C_{\alpha-1} - C_{-1} - N - C_{\alpha}$

Como la posición de los tres átomos de cada aminoácido es fija, esto nos deja con tres grados de libertad por cada enlace peptídico. Debido a las características del enlace peptídico que se forma entre aminoácidos el ángulo ω sólo puede tener dos valores $\omega = 0°$, si el enlace es de tipo cis o $\omega = 180°$, si el enlace es trans. Esto se reduce aun más ya que en la naturaleza el 99 % de las proteínas tienen enlaces peptídicos de tipo trans.

De esta forma las dimensiones del espacio necesario para representar una proteína de n aminoácidos de longitud es $\mathbb{R}^{2(n-1)}[0, 2\pi)$ y $\mathbb{N}^n$. Esto equivalen a dos ángulos diedros para cada enlace peptídico $\mathbb{R}^2[0, 2\pi)$ y se tiene un enlace peptídico para cada pareja de aminoácidos $n-1$ y un entero para representar la cadena lateral de cada aminoácido.

3.2. Estructura inicial

La estructura con la que se inicia la búsqueda se puede suministrar de dos formas. La más sencilla es mediante un archivo PDB con una estructura ya definida, pero esto sólo es bueno si se quiere refinar algún resultado previo. Para el problema real es obvio que no se tendrá un estructura por cristalografía, ya que esto es lo que se desea buscar. En estos casos lo único que se conoce de la proteína a plegar es su secuencia. Utilizando un archivo FASTA[1] se introduce la secuencia

[1]Este es un formato a base de texto que se utiliza para representar secuencias de ADN o proteínas. En la primera linea se encuentra el identificador de la secuencia. En las lineas siguientes se representa la secuencia deseada utilizando la nomenclatura de un solo carácter, cada una de estas lineas no puede exeder los 80 caracteres.

de aminoácidos de la estructura problema. Posteriormente un programa genera un modelo uniendo aminoácidos individuales extraídos de imágenes de cristalografía; para ello, se conoce un ejemplar para cada uno de los rotámeros de cada aminoácido [24]. Para reducir el número de variables se tomó sólo el rotámero más probable para cada aminoácido. En el caso de las funciones de energía que sólo dependen del esqueleto, este paso se puede eliminar y sólo trabajar con el esqueleto, ya que las cadenas laterales se pueden agregar al modelo final.

Los aminoácidos son unidos simulando un enlace peptídico: con la distancia promedio que se observa en cristalografía entre el carbono y el nitrógeno del siguiente aminoácido [7] y en enlace trans ($\omega = 180°$). Los dos ángulos diedros restantes son tratados de dos formas distintas. En el caso sencillo simplemente se les asigna valores aleatorios a todos colocando así la estructura inicial en un punto al azar en el espacio de búsqueda.

Con un enfoque más ambicioso (substitución) se busca copiar segmentos enteros de alguna de las estructuras objetivo; esperando seleccionar elementos de estructura secundaria y así se espera que el punto de inicio comience más cercano al mínimo local. Este método se describe con más detalle adelante.

En ambos casos, el oxígeno es colocado en su posición de acuerdo a la orientación del aminoácido que lo precede.

La substitución de la estructura inicial se lleva a cabo tomando al azar fragmentos de diferentes tamaños y varias partes de alguna de las estructuras objetivo, estos se copian directamente a la secuencia inicial, copiando los ángulos diedros, en el orden en el que se van escogiendo. Entre cada segmento que se copia se deja un espacio libre donde los ángulos se colocan al azar. Los fragmentos copiados no sólo comienzan de esa forma en el plegado inicial sino que se conservan a lo largo de toda la búsqueda. Así, los elementos de estructura secundaria que se llegan a copiar no se pierden mientras avanza la búsqueda. Como varios de los ángulos diedros entre aminoácidos se fijan, el espacio de búsqueda se reduce del original. Así no sólo se logra acercar el punto inicial al mínimo global sino que a la vez se reduce el espacio de búsqueda.

Un tercer método que se consideró fue el restringir el espacio de búsqueda a una lista que tenga los ángulos diedros que se observan en imágenes cristalográficas.

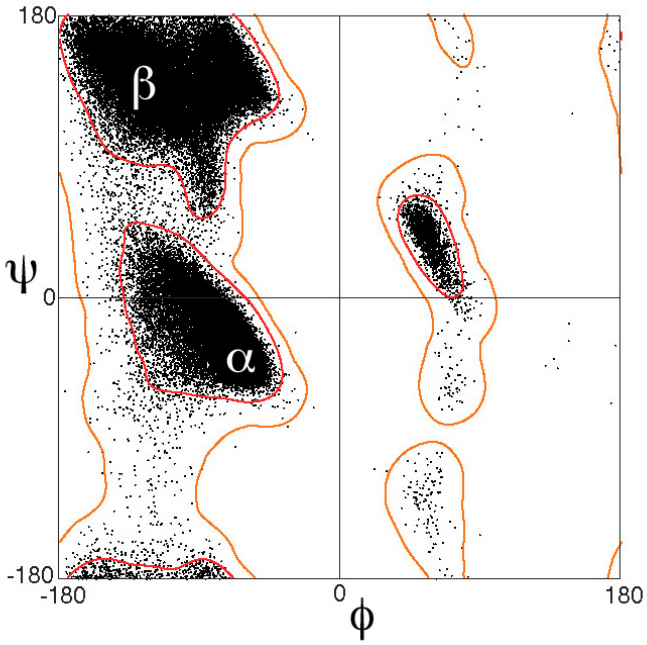


Figura 3.4. Diagrama de los valores permitidos para los ángulos diedros.

3.3. Función de energía

Una vez que se tiene una representación numérica de plegado de las proteínas se le debe asignar una función de energía. Para esto es necesario definir claramente el objetivo de la búsqueda, y luego elegir una función que lo establezca como el mínimo global. Normalmente se conocen propiedades importantes del objetivo y la función se elige de acuerdo a esto.

En este problema en particular, el objetivo es obtener la estructura plegada de una proteína, la cual se desconoce, pero hay varias características que se sabe que debe cumplir. Dado que es una estructura estable dentro de la célula, se supone que tiene que ser una conformación que maximice los contactos débiles.

Otro factor que se puede tomar en cuenta son las interacciones hidrofóbicas, buscando aumentar la exposición de moléculas hidrofílcas y minimizar la de las hidrofóbicas.

También se puede aprovechar la información que se ha obtenido por cristalografía, tomando una o varias proteínas que se sabe que tiene una estructura similar a la secuencia problema y utilizar un criterio de similitud entre ellas.

La motivación inicial de este trabajo era utilizar los vectores RCC (sección 3.3.4 3.5) para comparar dos estructuras, por lo que se usaron métodos de similitud estructural como control. Una

vez que se determinó que los RCC no eran útiles se continuó todo el trabajo utilizando comparación estructural.

3.3.1. Similitud en secuencia

Ya existen muchos métodos para evaluar la similitud entre dos estructuras. Los más sencillos simplifican el problema sólo buscando la similitud entre los esqueletos de las proteínas. La mayoría comienzan intentando alinear una estructura sobre la otra. Una vez hecho esto, se le asigna un valor a la superposición de éstas. Cada uno de estos pasos tienes sus complicaciones.

El alinear dos figuras tridimensionales es un problema complejo. Para las proteínas, en muchos casos se agiliza el alineamiento estructural realizando primero un alineamiento en secuencia, es decir, se buscan las similitudes entre las cadenas que forma las proteínas y basados en esa similitud se realiza el alineamiento estructural. Pero no siempre es útil usar como base la similitud en secuencia.

Se conocen numerosos ejemplos de estructuras extremadamente similares en secuencia (más del 90 %) pero cuyos plegados finales son completamente diferentes (menor al 20 %). Aun así, todavía existe la creencia de que la similitud en secuencia siempre implica un plegado igual. Aunque esto es cierto, en la mayoría de los casos es preferible, en la medida de lo posible, evitar o al menos tener la opción de no usar los alineamientos en secuencia, para así abarcar una panorama más general.

Existen menos métodos de alineamiento independientes de secuencia, éstos sólo utilizan la geometría de las estructuras para alinearlas y así poder evaluar su similitud. Por esto suelen ser más lentos que los basados en secuencia.

El problema de asignar un puntaje no se resuelve simplemente con el alineamiento, no hay una única forma de evaluar el parecido de dos figuras tridimensionales. Normalmente se utiliza la *root mean square deviation* (RMSD), entre los carbonos alfa de la proteína objetivo con el modelo.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_{ij}^2}.$$

donde d_{ij} es la distancia o diferencia entre 2 puntos que se comparan.

Para esto es necesario aparear los carbonos alfa de las dos cadenas, esto sólo es sencillo para las zonas donde el alineamiento emparejó las dos cadenas de una forma similar

Para el resto de la cadena no es claro cuál es la pareja “correcta” de cada carbono alfa. No

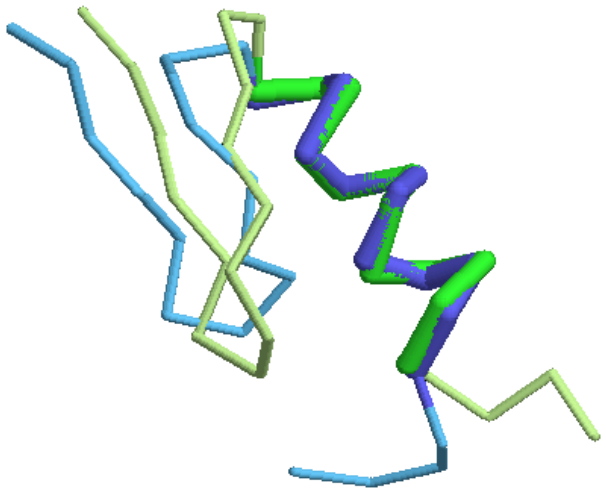


Figura 3.5. Alineamiento local de dos cadenas de proteína. La parte gruesas representa la zona que se logro alinear, el resto es descartado.

siempre es correcto simplemente continuar las parejas una vez que se separan ya que puede que algún otro carbono alfa se encuentre más cerca. Por esto, normalmente, la puntuación que se le asigna a un alineamiento sólo dependen de las zonas bien emparejadas, es decir, es una puntuación de similitud local. Estas funciones "descartan" las partes que no lograron alinear y sólo evalúan las que sí se logró, por esto son útiles para conocer la similitud de estructuras que cuentan con algunas zonas muy similares pero el resto difiere.

Este es un buen ejemplo de una función de energía que entorpece la optimización, ya que una similitud local no refleja el parecido real de dos estructuras. Si se quiere utilizar un puntuación de tipo RMSD, es necesario extender los métodos tradicionales de forma que siempre se considere la totalidad de la cadena en el cálculo. En la figura 3.5 se muestra un ejemplo de un alineamiento local.

3.3.2. Smith Waterman

Temple F. Smith y Michael S. Waterman en 1981 propusieron un método de alineamiento local en secuencia, es decir, busca alinear zonas del mayor tamaño posible en lugar de toda la secuencia [15].

Para hacer esto se coloca una de las secuencias a un lado de los renglones de una matriz y la otra encima de las columnas, luego se inicializan el primer renglón y columna con ceros.

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \mathbf{2} & 1 & 2 & 1 & 2 & 1 & 1 & 2 \\ G & 0 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 1 & \mathbf{3} & 2 & 3 & 2 & 3 & 2 & 2 \\ A & 0 & 2 & 2 & \mathbf{5} & 4 & 5 & 4 & 4 & 4 \\ C & 0 & 1 & 4 & 4 & \mathbf{7} & 6 & 7 & 6 & 6 \\ A & 0 & 2 & 3 & 6 & 6 & \mathbf{9} & 8 & 8 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & \mathbf{11} & \mathbf{10} & 10 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & \mathbf{12} \end{pmatrix} \quad T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \mathbf{\uparrow} & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \mathbf{\nwarrow} \end{pmatrix}$$

Figura 3.6. Matriz para alineamietno Smith Waterman.

La matriz se va llenando utilizando una función de puntuación. Esta utiliza los valores adyacentes anteriores ($[i-1, j], [i, j-1], [i-1, j-1]$) y una matriz de puntuación en la que se asigna un valor para cada pareja de aminoácidos (de acuerdo a algunas propiedades o similitudes entre éstos); la función toma el valor máximo de estas 3 opciones o 0 si todos resultan ser negativos.

Una vez llenada toda la matriz, partiendo del valor máximo, se traza un camino hacia arriba y hacia la izquierda, avanzando siempre por los valores más altos. Las flechas diagonales indican un alineamiento, las horizontales una inserción y las transversales una deleción. En la figura 3.6 se muestra un ejemplo de este procedimiento. Con el conocimiento de las parejas alineadas se usa esta información para acelerar el alineamiento tridimensional, ya que éste simplemente se reduce a disminuir la distancia de todas los aminoácidos apareados. Este método sirve como una buena función de energía siempre y cuando las dos estructuras sean suficientemente parecidas en secuencia.

3.3.3. SPalignNS

El método SPalignNS [20] es una modificación independiente de secuencia del método SPalign [19] utilizando una solución general del problema *Linear Sum Assignment Problem* (LSAP) para matrices asimétricas. El programa que hace el alineamiento es libre y fue modificado ligeramente [32] para calcular el RMSD de la totalidad de las moléculas y no solo hacer el cálculo local.

3.3.4. RCC

También existe otra opción independiente de un alineamiento en secuencia para calcular la similitud entre dos estructuras. Si se tiene una función que identifique cualquier estructura con un punto en $\mathbb{R}^n$, el problema se simplifica drásticamente porque sólo es necesario calcular una distancia eu-

clidiana entre dos puntos para evaluar su parecido. En este caso, el problema radica en construir una función que refleje acertadamente las características de la forma de una proteína utilizando sólo n valores.

Existe una función que hace esto: los *Residue Cluster Classes* (RCC) [11] son vectores en $\mathbb{N}^{26}$ que enumeran el tipo y la cantidad de cúmulos que se forman en una proteína; está demostrado que estos logran dividir el espacio de plegados de las proteínas en clases similares.

3.4. Refinamiento

El refinamiento consistió en una versión modificada del recocido simulado para la cual sólo se aceptan las soluciones que disminuyan la energía actual. Este método modificado se repiten varias veces utilizando ciclos cortos donde cada uno toma la mejor solución del anterior. Con esto se busca simplemente acercarse al mínimo más cercano. Así, se mejorar una solución que se espera que ya esté muy cerca del objetivo.

3.5. Resultados previos

Aunque los ángulos diedros permitidos para las conformaciones nativas se conocen, no se utilizó esto durante ninguna búsqueda. El restringir el espacio de búsqueda a solo algunos valores entorpece la exploración libre del espacio y puede hacer ciertos puntos inaccesibles.

En primera instancia, los RCC toman en cuenta las cadenas laterales. Por esto el método es extremadamente lento (como se ve más adelante en la tabla 4.1 y la gráfica 4.1). Se podría resolver esto previamente generando una base de datos para la cuál previamente se calcularon sus vectores RCC. Utilizar un sistema eficiente de búsqueda, prácticamente se eliminaría el tiempo de cálculo. Éste diseño valdría la pena si el método presentase resultados prometedores.

Pero, al hacer varias pruebas utilizando los RCC, se observó que el espacio que generan refleja de forma muy precisa cualquier ligera diferencia entre dos estructuras muy similares. Estructuras con una similitud en RMSD menor a 1×10^{-5} resultan tener una distancia de más de 20 unidades en el espacio de RCC. A primera vista, esto puede parecer una ventaja; pero esta sensibilidad se pierde por completo cuando se trata de estructuras muy distintas. En promedio, cualquier estructura generada al azar resulta tener una distancia a cualquier otra de por lo menos 50 unidades y ésta

diferencia no disminuye hasta que las estructuras comienzan a tener similitudes muy importantes.

Por todas estas razones, se abandonó por completo la utilización de los RCC y se continuó utilizando métodos de comparación estructural.

Capítulo 4

Resultados

Para las pruebas se eligió un conjunto representativo con proteínas de distintas longitudes y clases de acuerdo a la clasificación CATH [29]; también se utilizaron los identificadores de CATH para referirse a las distintas proteínas. En todos los casos las pruebas intentaron reproducir la proteína objetivo partiendo de la secuencia de la misma. Es decir, la función objetivo fue la diferencia entre el modelo actual generado de la secuencia y la estructura cristalográfica conocida.

4.1. Pruebas de velocidad para funciones de energía

Primero se evaluaron distintas funciones de energía sobre todo el conjunto de pruebas (tabla 4.1). Para todos los casos se comparó cada proteína consigo misma.

Excepto el primero (RCC) todos ellos son métodos de alineamiento y de estos sólo el último (SPAlignNS) no utiliza alineamiento secuencial. Los métodos de alineamiento CE [31] y FATCAT [30] se basan en alineamiento local y sólo se usaron como referencias del tiempo de ejecución.

En la figura 4.3 se muestra el resultado de una búsqueda utilizando el método de alineamiento CE. Al igual que en la figura 3.5 el usar alineamiento local no da buenos resultados; sólo se alinea la zona del medio mientras que los dos extremos se ignoran por completo. En la gráfica 4.1 se ve cómo crece el tiempo de cálculo mientras va aumentando la longitud de las proteínas, y en la gráfica 4.2 se muestra el mismo comportamiento pero sólo comparando los dos métodos más rápidos. Todas las funciones presentan el mismo tipo de crecimiento. Por lo que, para cualquier longitud, el método más rápido es el de Smith Waterman.

Identificador CATH	Longitud	RCC	FATCAT	CE	Smith Waterman	SPalignNS
1b9m	255	214.33	14.89	3.09	0.67	1.73
1eex	551	994.17	80.77	17.49	1.13	3.88
1f8n	818	240.40	151.58	43.36	1.96	6.12
1n7h	331	387.00	28.78	5.41	0.80	2.31
1orn	214	154.75	22.48	2.76	0.71	1.57
1pda	296	285.18	18.28	5.28	0.63	1.84
1thg	543	998.40	69.68	16.04	1.21	4.03
1tw9	191	136.19	13.55	2.20	0.50	1.35
2ad6	571	108.45	80.58	18.90	1.34	3.72
2b1y	98	39.84	3.03	0.59	0.41	0.90
2fcp	705	169.20	115.83	34.70	1.61	4.90
2hox	425	653.33	42.87	9.72	1.00	2.83
2j98	106	43.30	3.67	0.60	0.43	1.00
2p1g	230	183.41	13.88	3.00	0.54	1.41
3ag6	283	279.58	21.27	4.29	0.64	1.90
3gvk	644	146.81	122.94	26.63	1.40	4.47
3ppm	546	987.68	81.99	16.83	1.22	3.68
3q9k	594	125.17	82.87	20.50	1.24	4.02
3riq	541	989.63	109.91	18.17	1.28	3.81
3s8f	554	107.57	167.42	18.12	1.26	4.13

Tabla 4.1. Tiempo de ejecución (segundos) de un ciclo de búsqueda para distintas funciones de energía en todo el conjunto de proteínas de prueba.

En la gráfica 4.1 se ve cómo el tiempo de ejecución calculando los RCC crece extremadamente rápido y opaca al resto de las funciones de energía. Esta es una de las razones por la que no se decidió continuar utilizando esta función.

4.2. Pruebas de eficiencia

Se eligió la función de energía más rápida para realizar pruebas de eficiencia sobre todo el conjunto.

Cada búsqueda se ejecutó con los siguientes parámetros:

- Un total de 3000 ciclos de búsqueda, disminuyendo la temperatura cada 5 ciclos.
- Temperatura inicial de 2.5 deteniéndose en 1.
- La función de energía se multiplicó por 100 para que aumentar el rango de aceptación.

Relación entre el tiempo de ejecución y la longitud de la proteína.

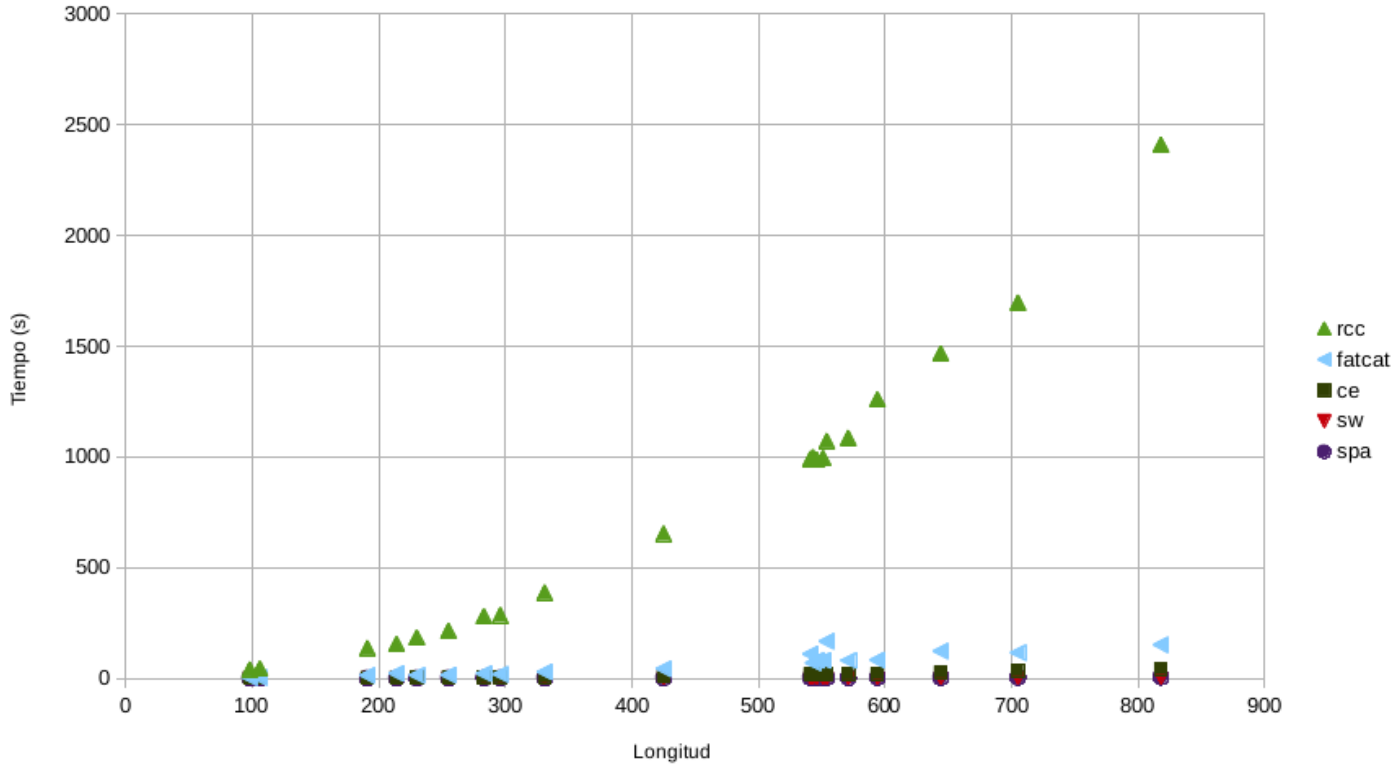


Figura 4.1. Gráfica de la tabla 4.1 Para todas las funciones de energía el aumento en el tiempo es proporcional a la longitud de la proteína.

Relación entre tiempo de ejecución y la longitud de la proteína.

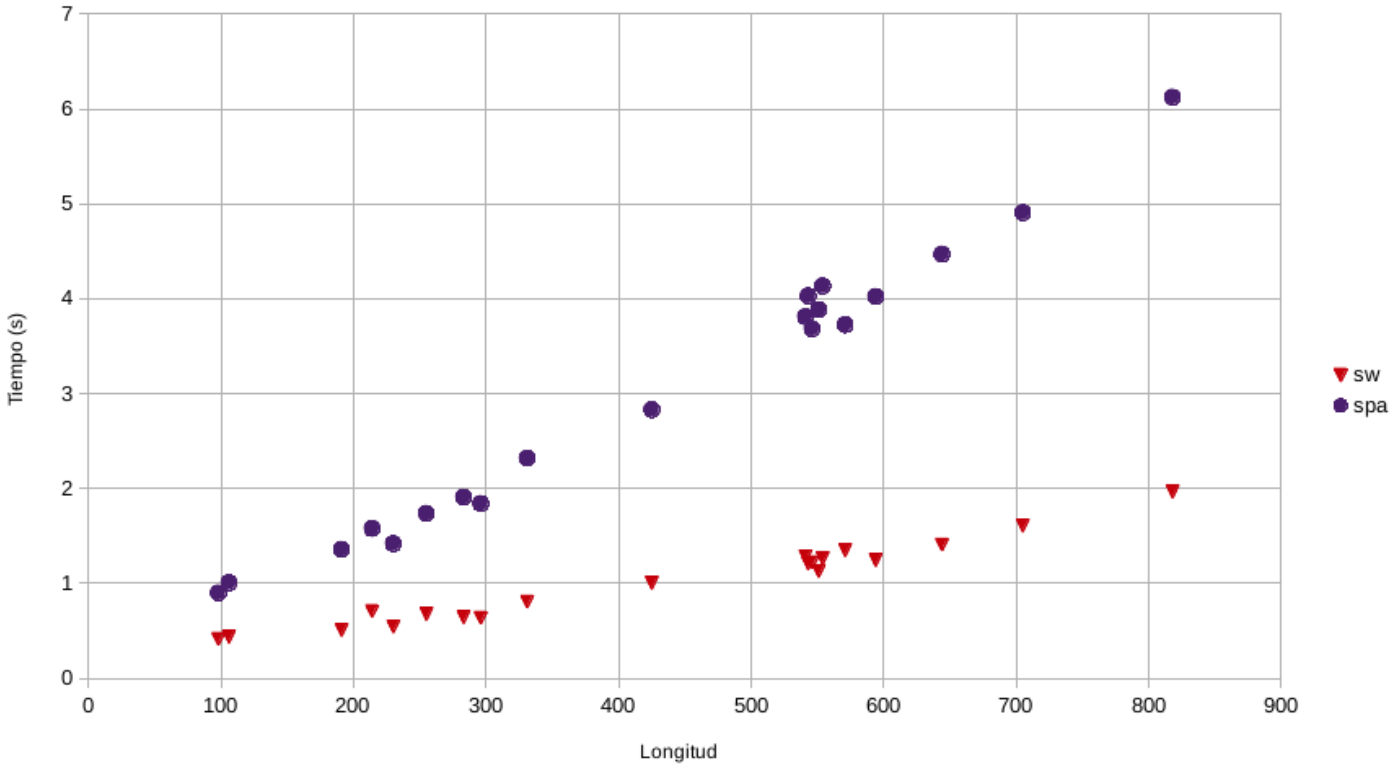


Figura 4.2. Desempeño de los dos métodos que se utilizaron para calcular la energía. El método Smith-Waterman (sw) es más rápido que el SPalignNS (spa).

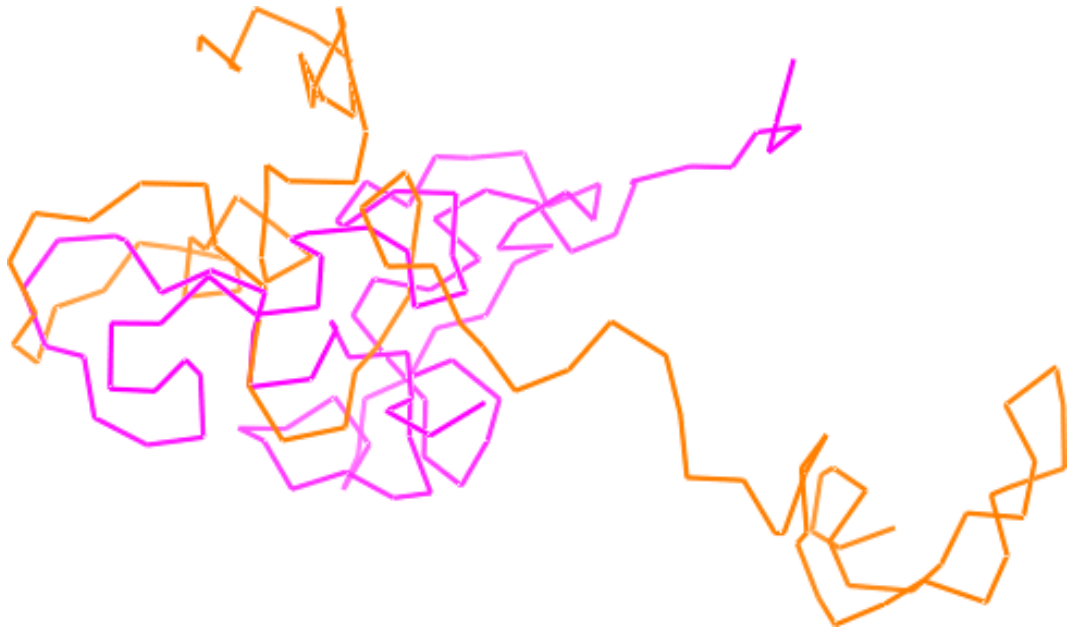


Figura 4.3. Resultado de una optimización utilizando el método de alineamiento CE. En naranja se muestre en modelo y en morado el objetivo. Los alineamientos locales sólo optimizan una zona, en este caso el centro, e ignoran el resto de la estructura.

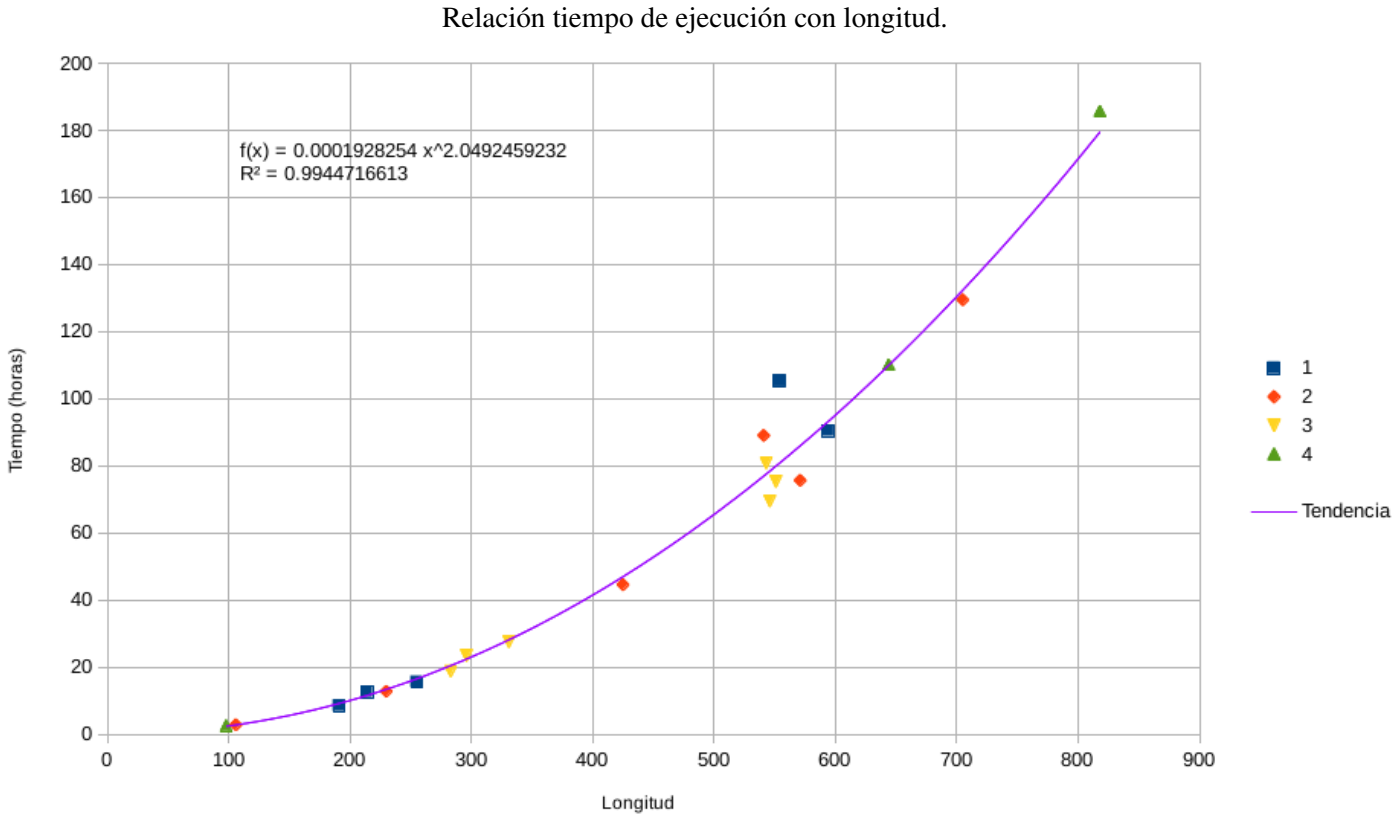


Figura 4.4. Los números a la derecha indican la clase CATH a la que pertenece la proteína. La distribución se ajusta muy bien a una función cuadrática por lo que es fácil concluir que la relación entre el tiempo de optimización y la longitud de la proteína tienen una relación cuadrada.

- El margen de variación para cada ángulo diedro comenzó en 9 y fue disminuyendo junto con la temperatura hasta llegar a 0.001.
- Se ejecutaron 1000 de estas búsquedas en paralelo, todas con estructuras iniciales distintas elegidas al azar modificando todos los ángulos diedros.

Todos estos parámetros se eligieron por medio de prueba y error tras haber explorado como afectaban el desempeño de la búsqueda.

Los resultados se muestran en la tabla 4.2 y las gráficas 4.4, 4.5 y 4.6. Y en las figuras 4.7b 4.8c se pueden ver los resultados gráficos de algunas de las búsquedas.

Se observa que la clase a la que pertenece cada proteína no tiene ninguna relación que el desempeño del a búsqueda, en todos los resultados son independientes a la clase estructural de la proteína y los resultados finales sólo son afectados por la longitud. En la gráfica 4.5 se ve que el tiempo de ejecución crece cuadráticamente con la longitud y en la gráfica 4.6 que el puntaje final tiene una

Identificador CATH	Longitud	Tiempo	RMSD
Clase 1			
1tw9 A02	191	8.45	6.45
1orn A02	214	12.56	6.63
1b9m A01	255	15.54	8.28
3s8f A00	554	105.31	12.99
3q9k A00	594	90.44	13.16
Clase 2			
2j98 A01	106	2.76	5.41
2p1g B02	230	12.74	7.52
2hox A01	425	44.54	11.34
3riq A00	541	89.04	13.07
2fcp A02	563	129.47	18.15
2ad6 A00	571	75.62	13.21
Clase 3			
3ag6 A02	283	18.73	8.63
1pda A02	296	23.35	9.32
1n7h A01	331	27.48	9.83
1eex A00	551	75.26	12.57
1thg A00	543	80.78	12.03
3ppm A00	546	69.45	11.80
Clase 4			
2b1y A00	98	2.47	5.45
3gvk A04	644	110.09	15.38
1f8n A02	818	185.70	15.28

Tabla 4.2. Pruebas de eficiencia utilizando la función de energía Smith-Waterman para todo el conjunto de pruebas.

El tiempo se reporta en horas. Las 4 clases corresponde a la clasificación CATH: plegado alfa, beta, alfa+beta y escasa estructura secundaria.
En la tabla se reporta la menor energía (RMSD) obtenida de entre las 1000 búsquedas simultaneas.

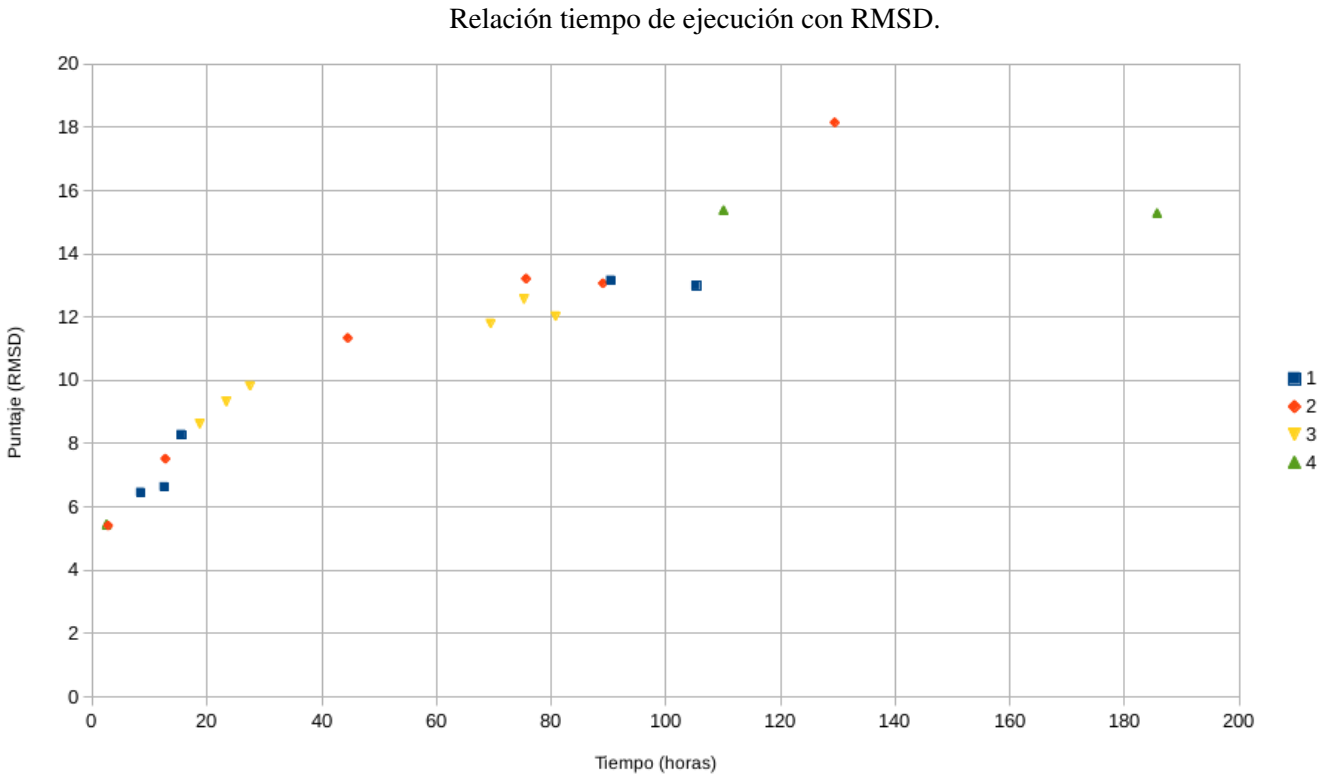


Figura 4.5. La representación de la derecha es igual a la de la tabla 4.4. Aquí aunque la relación entre el puntaje final y el tiempo de ejecución es creciente no hay ninguna función simple a la que se ajuste la distribución.

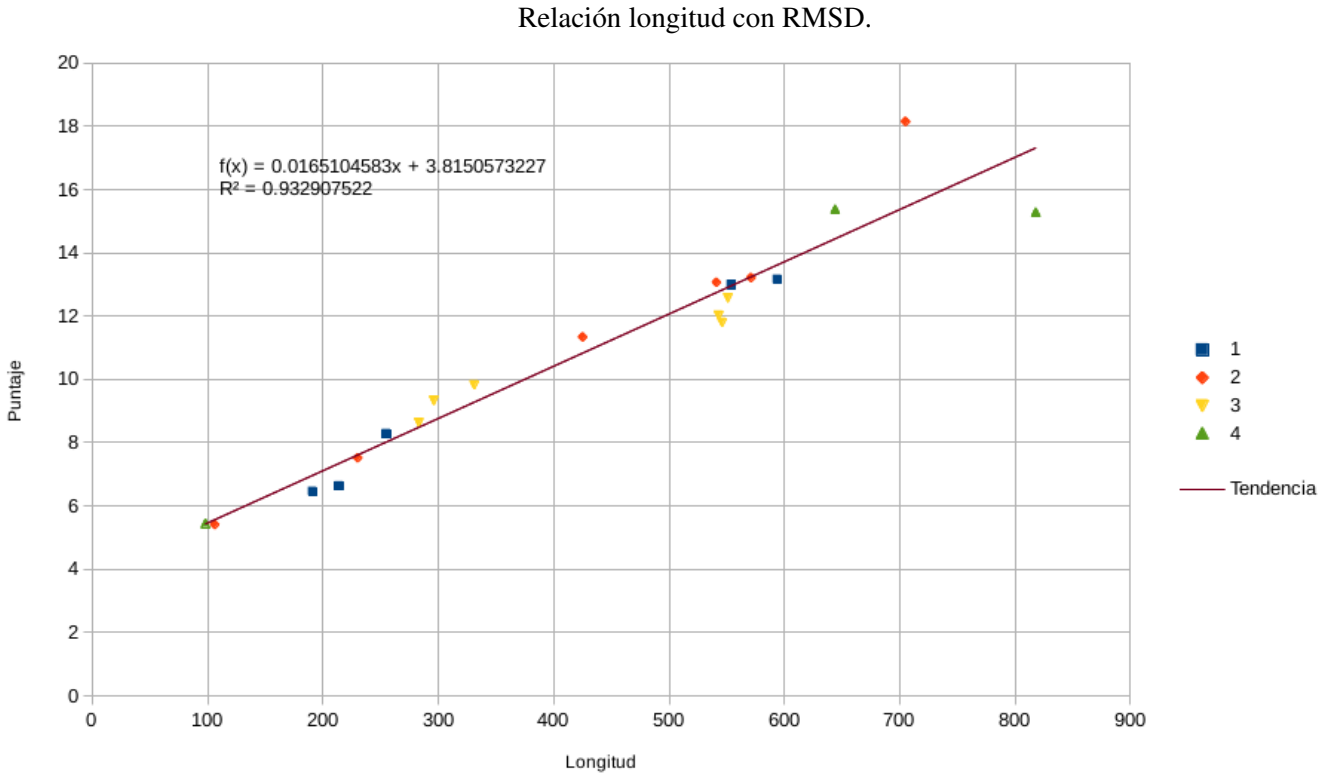
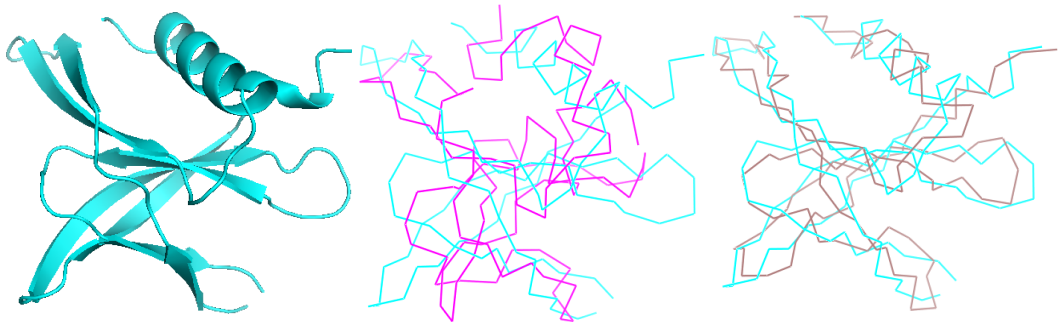


Figura 4.6. A la derecha se usa la misa representación de la tabla 4.4. Se puede ver que la longitud de la proteína a optimizar afecta el puntaje final que se obtiene. La relación parece ser lineal.



(a) Estructura de la proteína 2j98 A.

(b) Modelo obtenido con búsqueda (en magenta).

(c) Modelo tras refinamiento (en naranja).

Figura 4.7. Comparación de los modelos obtenidos para la proteína 2j98 A

(a) Estructura de la proteína 1n7h A.

(b) Modelo obtenido con búsqueda.

(c) Comparación del modelo con 1n7h.

Figura 4.8. Modelo obtenido para la proteína 1n7h A

relación lineal con la longitud.

Para estructuras con longitudes mayores a 500 aminoácidos el tiempo que tarda rebasa los 8 días. Se esperaba algo así por los resultados de las pruebas de velocidad; además las modificaciones al azar que también se realizan en cada paso del método crecen al aumentar la longitud.

En la gráfica 4.6 se observa una relación (más o menos lineal) entre la longitud y el puntaje RMSD. Es posible que esto se deba a que una mayor longitud implica un espacio de búsqueda más grande que a su vez dificulta la ubicación de mínimos. Aunque se obtienen valores por debajo de 10 para todas las estructuras cortas (longitud menor a los 300 aminoácidos), se puede ver en la figura 4.7b que su similitud con la estructura nativa es pobre. Se utilizó el refinamiento para intentar mejorar estos modelos. Es de esperarse que éstos ya se encuentren cerca de un grán mínimo global, por lo que solo es necesario acercarlos a éste.

Se intentaron refinar los mejores resultados para cada clase utilizando los siguientes parámetros:

- Un total de 1000 ciclos de búsqueda, disminuyendo la temperatura cada ciclo.
- Temperatura inicial de 4 deteniéndose en 1.
- La función de energía se multiplicó por 100.
- Se realizaron 15 ciclos sucesivos cada uno intentando mejorar el resultado anterior.
- Variación de ángulos comenzando en 0.01 y terminando en 0.001; la variación inicial fue diminuyendo con cada ciclo de manera lineal hasta terminar en 0.0001 y 0.000001, respectivamente.

Clase	Identificador CATH	Longitud	Score Inicial	Score Refinado	Porcentaje reducido
1	1tw9 A02	191	6.45	4.772	26.09
2	2j98 A01	106	5.41	3.176	41.33
3	3ag6 A02	283	8.63	5.967	30.85
4	2b1y A00	98	5.45	3.276	39.02

Tabla 4.3. Refinamiento de un resultado para cada clase.

Como los puntajes iniciales no son iguales para las 4 clases se utilizó un porcentaje de mejora con respecto a su resultado inicial para comparar la mejora al refinar.

Estos cálculos se ejecutaron en tan sólo unos minutos. Los resultados se muestran en la tabla 4.3.

Aunque sólo se tiene una muestra para cada clase, los resultados de la tabla 4.3 coinciden con lo observado durante las pruebas para determinar los parámetros óptimos de búsqueda: Las hélices alfa resultan más difíciles para el método. Al refinar la estructura, la mejora para la clase 1 (principalmente alfa) es la menor, seguida por la clase 3 (alfa+beta)

Este refinamiento se puede continuar indefinidamente pero la mejora lograda en cada paso es cada vez menor. Para demostrar esto se tomó el modelo obtenido para la proteína 2j98, que tuvo el mayor porcentaje de reducción, y se refinó de la misma forma otras 5 veces obteniendo un valor RMSD final de 2.594. Este modelo se muestra en la figura 4.7c.

4.3. Substitución

Para evaluar la utilidad de la substitución se hicieron dos pruebas utilizando la técnica de substitución. Se eligió una estructura con buenos resultados en las pruebas de eficiencia (puntaje RMSD de 6.63) y otra con resultados menos favorables (puntaje RMSD de 11.34). Los fragmentos conservados fueron tomados al azar de la estructura objetivo, variando la cantidad y la longitud. En estas pruebas los fragmentos se colocaban en la misma posición en la que se encontraban en el objetivo. Los parámetros de búsqueda fueron los mismos que para los cálculos de la tabla 4.3. Los resultados se muestran en la tabla 4.4. Aunque los mejores resultados están dominados por fragmentaciones grandes (50 % de la molécula substituida y conservada), entre los mejores 100, más de 15 estaban por debajo de 20 %. Por ello, la substitución de grandes fragmentos parece ayudar a la búsqueda pero no es determinante.

Tabla 4.4. Búsuqeda con Substitutor

ID Proteína	Longitud	Tiempo	Tiempo Sub	Score	Score Sub
1orn A	214	12.05	12.56	6.63	5.237
2hox A	425	48.97	44.54	11.34	8.377

Se compara el tiempo de ejecución y el puntaje final de las dos estrategias: substitución y variando la estructura libremente.

Para la molécula chica no se observa una diferencia importante ni en tiempo ni en puntaje. Pero para la de mayor longitud el método es 4 horas más veloz y obtiene un resultado 26 % menor.

4.4. Alineamiento independiente de secuencia

Por último, se hizo un cálculo usando la función de energía SPAlignNS, que alinea las estructuras independientemente de la secuencia.

Los parámetros utilizados fueron similares a los de la tabla 4.2.

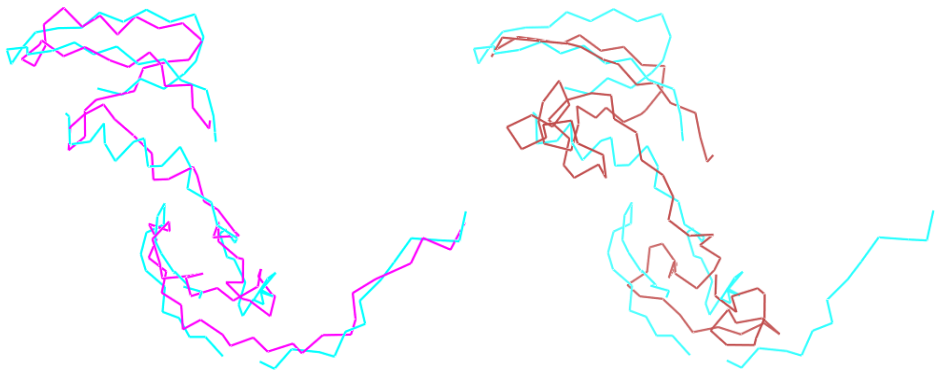
- Técnica de substitución.
- 3000 ciclos de búsqueda, disminuyendo la temperatura cada 5 ciclos.
- Temperatura inicial de 2.5 deteniéndose en 1.
- Función de energía SPAlignNS modificada para calcular RMSD global.
- Variación de ángulos comenzando en 9.0 y terminando en 0.001.
- 1000 búsquedas en paralelo con semillas iniciales y substituciones distintas.

Posteriormente, se realizaron dos ciclos de refinamiento sobre el mejor resultado, uno respetando la substitución y otro moviendo toda la estructura libremente. Para ambos se utilizaron los siguientes parámetros:

- 1000 ciclos de búsqueda, disminuyendo la temperatura cada ciclo.
- Temperatura inicial de 4 deteniéndose en 1.
- Función de energía SPAlignNS.

Tabla 4.5. Búsqueda con SPAlignNS comparada con Smith Waterman

ID Proteína	Longitud				
2b1y A	98				
Método	Tiempo	Score	Promedio	Refinamiento	Refinamiento con Substitución
Smith Waterman	12.05	5.453	7.13	2.887	
SPAlignNS	35.57	5.247	7.227	4.177	4.993



(a) Alineamiento del modelo obtenidido con Smith Waterman (magenta) con 2b1y.

(b) Alineamiento del modelo obtenidido con SPAlignNS (naranja) con 2b1y.

Figura 4.9. Comparación de modelos obtenidos con distintos métodos después de ser refinados libremente.

- Se realizaron 30 ciclos sucesivos cada uno intentando mejorar el resultado anterior.
- Variación de ángulos comenzando en 0.01 y terminando en 0.001, la variación inicial fue disminuyendo con cada ciclo de manera lineal hasta terminar en 0.001 y 0.00001 respectivamente.

Como control se hizo la misma búsqueda y refinamiento libre utilizando Smith Waterman como función de energía. Los resultados están en la tabla 4.5 y la figura 4.9. Aquí se ve la importancia que tiene la velocidad de cálculo para el método de alineamiento. La ligera diferencia que hay entre Smith Waterman y SPalignNS (0.41 y 0.9 segundos respectivamente) crece a casi el triple al realizar una búsqueda en paralelo.

La comparación de la eficiencia de una función de energía independiente de secuencia con una dependiente revele que el resultado obtenido es muy similar para los dos, con una diferencia de 0.206 en su puntaje RMSD. Los promedios de todos los resultados también resultan ser casi iguales, con una diferencia de 0.097. La diferencia más notable se obtiene al implementar el refinamiento. La superioridad del método dependiente de secuencia se puede ver tanto numéricamente, con una

diferencia de 1.29 en puntaje RMSD, como gráficamente.

Capítulo 5

Conclusiones

En esta tesis se estudiaron diversas variantes de un método para predecir las cadenas de proteínas por medio de técnicas computacionales.

Se puede afirmar que **el método sólo funciona para estructuras cortas** (menores de 300 aminoácidos) para las cuales el cálculo es bastante rápido y se obtienen resultados buenos por debajo de 10 en RMSD que posteriormente pueden ser refinados. Cabe recalcar que más del 80 % de los dominios de las proteínas conocidas tienen una longitud menor a 200 aminoácidos.

Las pruebas con las que se cuenta indican que las alfa hélices son más difíciles de refinar. Es probable que la causa de esto sea la aparición de muchos mínimos locales cuando la función de energía se aplica a una estructura de la clase 1.

Esto puede ser provocado por la forma de las hélices. Si en la primera aproximación el modelo simplemente pasa por la parte de en medio de la hélice del objetivo, al alinear, el valor RMSD será bajo. Pero la cantidad de aminoácidos que forman parte de la *hélice* del modelo serán mucho menores que los que forman la hélice del objetivo. Cuando se intenta refinar, el modelo debe "enroscarse" para así disminuir la distancia al objetivo. Al hacer esto, la cantidad de aminoácidos que forman parte de la *hélice* debe aumentar, para lo cual, se utilizan los aminoácidos de los extremos de la *hélice*. Dicho cambio modifica el resto del modelo aumentando el valor del RMSD.

Debido a esto, las hélices alfa se convierten en una zona de estrés para la cual es difícil disminuir el puntaje RMSD. En este caso puede ser que las soluciones con los valores RMSD más bajos no sean las que se encuentren cerca del mínimo global.

En cambio las otras estructuras, de la clase 2 y 4, pueden corregirse con una serie de pequeños

cambios locales que no afectan el resto de la estructura y alcanzar un valor menor.

Tras el proceso de refinamiento, el alineamiento independiente de secuencia (SPAlignNS) no presenta resultados mejores, a pesar de que en la búsqueda en paralelo se obtienen resultados similares al método dependiente de secuencia; su resultado final fue peor que el del alineamiento dependiente de secuencia. Cabe destacar que todas las pruebas se realizaron usando la secuencia de la proteína objetivo como secuencia inicial, esto implica que una técnica como Smith Waterman (que utiliza alineamiento en secuencia) logró un alineamiento perfecto por lo que todos los carbonos alfa estaban apareados. Si la secuencia inicial difiriese mucho del objetivo, se esperaría que el desempeño de Smith Waterman debiera de empeorar drásticamente mientras que SPAlignNS no se vería afectado.

La técnica de substitución parece mejorar el desempeño del método y ésta mejora es mayor para proteínas más largas. Se tiene una disminución del 26.1 % en valor RMSD con respecto al método con búsqueda libre. Aunque el tiempo aumentó para la proteína corta (1orn), el cambio fue de 4.2 %, lo cual es despreciable, acercándose a 0. El método se puede utilizar para aumentar el rango en el cual se obtienen resultados con valor RMSD menor a 10, sabiendo que la disminución en tiempo es poca: 2 % para la proteína 2hox.

La elección de un método basado en alineamiento como función de energía exige contar con un templado objetivo, lo cual implica que ya se tiene una idea de cómo será la estructura final. Gracias a la enorme cantidad de estructuras nativas que se conocen por cristalografía no es complicado tener uno o incluso varios templados para usar como objetivo.

La opción de utilizar más de un objetivo para la función de energía está implementada en el código pero se encuentra en etapa *beta*[1] en el programa que se hizo para este trabajo, pero no fue probada completamente. En ésta se calcula la diferencia RMSD del modelo a cada uno de los objetivos y se entrega el promedio de todas como energía final.

Si se conoce la clase a la que pertenece una secuencia desconocida, este método podría usarse para predecir la estructura final de la proteína utilizando varios objetivos de esa clase.

[1]Es decir, no se a probado lo suficiente para asegurar que esta libre de errores

Apéndices

Appendix A

Trigonometría compleja

A.1. Fórmula de Euler

$$e^{ix} = cosx + isenx$$

A.2. Ondas como vectores complejos

A veces es útil trabajar con exponenciales en lugar de con funciones trigonométricas, para esto se utiliza la parte real de la exponencial imaginaria:

$$Re\{e^{ix}\} = Re\{cosx + isenx\} = cosx$$

Además con esta representación también se facilitan los cambios de fase, multiplicar la exponencial por i recorre la fase $\frac{\pi}{2}$

$$Re\{ie^{ix}\} = Re\{icosx - senx\} = -senx$$

Se puede ver que, dado un complejo c, el ángulo que tiene en el plano complejo índica la fase de la onda y su norma la amplitud.

Además las suma vectorial en el plano complejo resulta ser la interferencia de ambas ondas

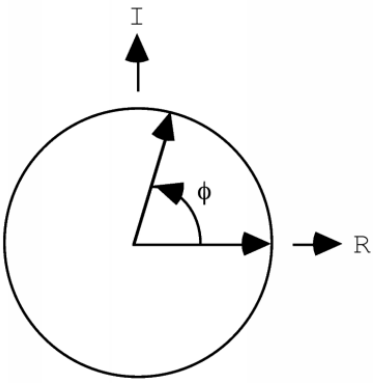


Figure A.1. Representación vectorial en espacio complejo de una onda.

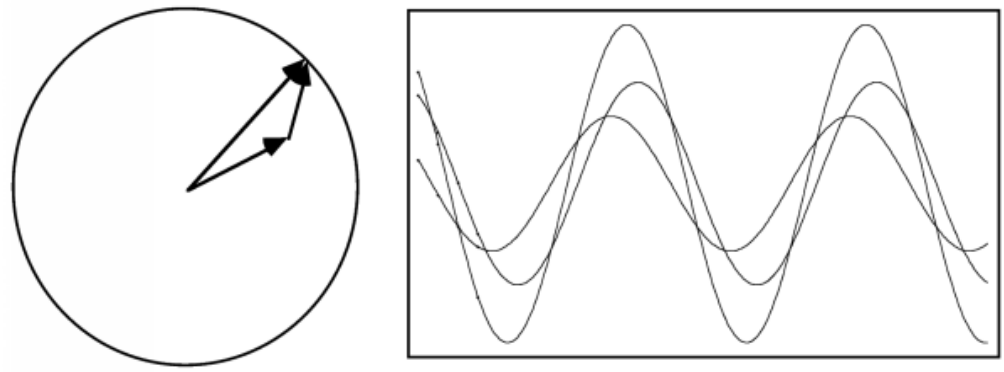


Figure A.2. Equivalencia entre suma vectorial e interferencia.

A.3. Series de Fourier

La expansión de Fourier de una función $f(x)$ en espacio complejo está dada por:

$$f(x) = \sum_{h=-\infty}^{\infty} C_h e^{i\frac{2\pi h}{a}x}$$

Appendix B

Planos de Bragg

Supongamos que se tiene un cristal con celda unitaria de paralelepípedo y lados $a \times b \times c$. Cada tripleta de enteros (h, k, l) define una familia infinita de planos paralelos conocidos como planos de Bragg de acuerdo a la ecuación:

$$\frac{xh}{a} + \frac{yk}{b} + \frac{zl}{c} = n.$$

con n cualquier entero. Los enteros (h, k, l) se conocen como los índices de Miller.

Por su definición, todos los planos están separados por la misma distancia d. Siempre cabe un número entero de planos dentro de la celda unitaria, además h planos cruzan su lado a, k su lado b y l su lado c (vease figura B.1).

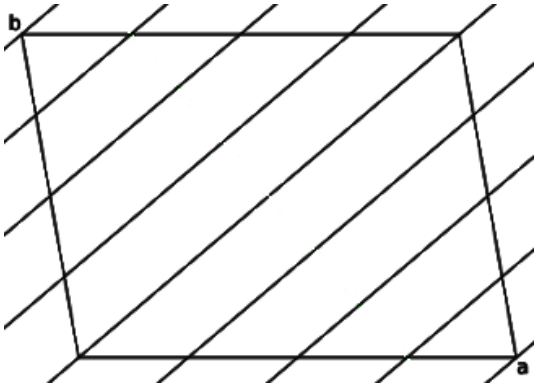


Figure B.1. Planos de Bragg definidos por los índices de Miller $(4 - 4)$.

Si la separación entre planos es d, la condición para que dos ondas incidentes con un ángulo θ

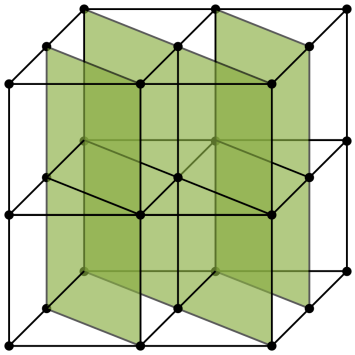


Figure B.2. Planos de Bragg definidos por los índices de Miller (011).

que se dispersan en dos planos interfieran constructivamente está dada por la ley de Bragg,

$$n\lambda = 2d \sin\theta \tag{B.1}$$

Appendix C

Interpretación del patrón de difracción

C.1. La esfera de Ewald

Para identificar a qué conjunto de planos de Bragg corresponde cada punto del patrón de interferencia se utiliza la llamada esfera de Ewald. Esta es una simple construcción geométrica con la que se puede conocer el ángulo de salida de los rayos difractados dado un conjunto de planos de Bragg.

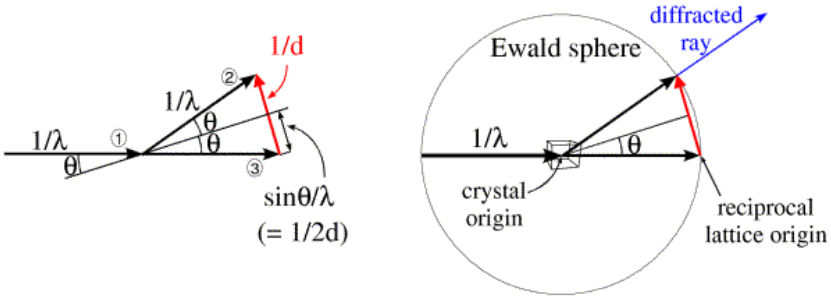


Figure C.1. Construcción de la esfera de Ewald.

Sea un rayo con longitud de onda λ que incide en el cristal con un ángulo θ. Si pensamos en el rayo incidente como un vector de magnitud $1/\lambda$, utilizando la ley de Bragg, se puede obtener que la diferencia entre el vector reflejado y el vector proyectado es $1/d$ con d la separación entre los planos de Bragg, como se ve en la figura C.1. Se sigue entonces que midiendo la distancia entre

los máximos de intensidad del patrón de difracción se puede conocer la separación de los planos de Bragg.

C.2. Contribución de cada punto

Para facilitar la explicación se va a tratar el problema para un cristal en dos dimensiones. Esta idea luego puede ser generalizada intuitivamente a tres dimensiones.

Dado un cristal simple (cúbico simple) y un conjunto de planos de Bragg con separación d, utilizando la ley de Bragg se sabe a que ángulo θ_0 se debe iluminar el cristal con un rayo de longitud de onda λ para obtener un máximo como salida.

El máximo es provocado por la interferencia constructiva. Para esto, la diferencia de fase entre el rayo reflejado en el primer plano y el reflejado en el segundo debe ser un múltiplo n de la longitud de onda.

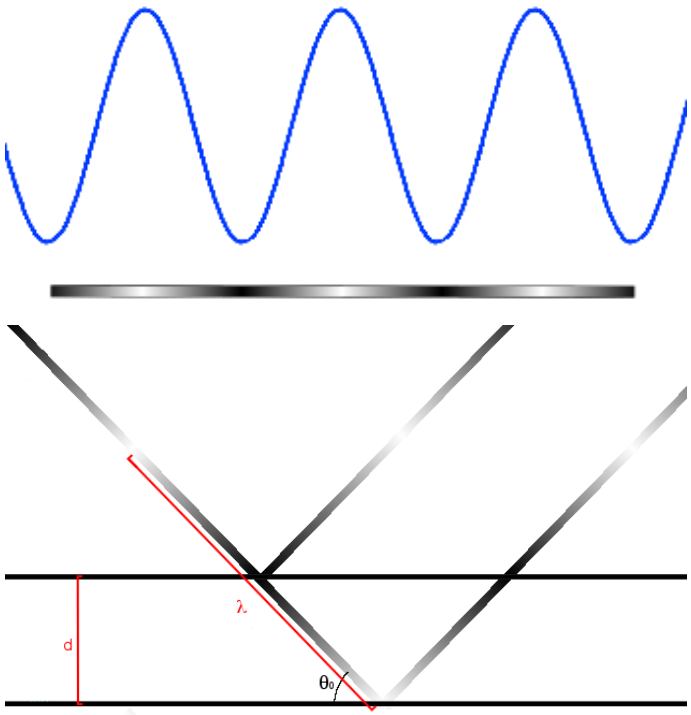
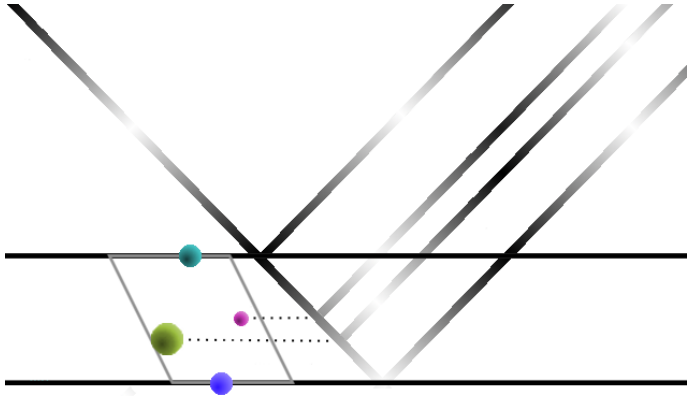


Figure C.2. Representación gráfica de la ley de Bragg.

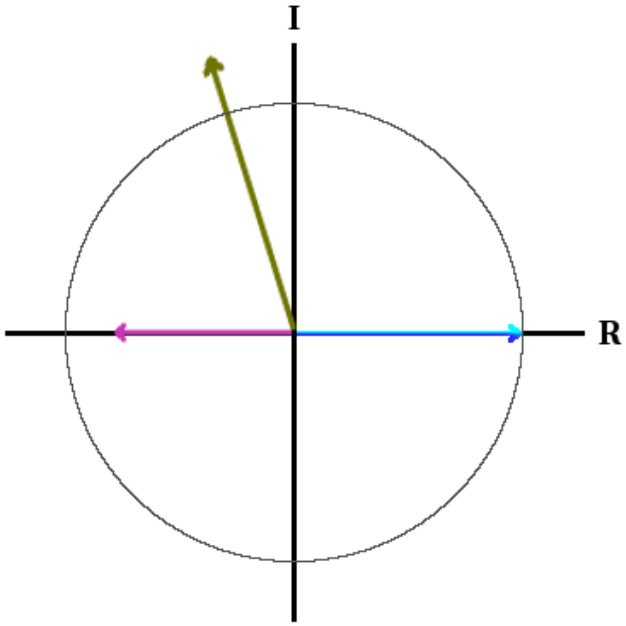
Pero si se tiene un cristal complejo (es decir con un patrón no ordenado en cada celda unitaria i.e. proteínas) la onda es reflejada en todos los átomos dentro de la celda unitaria. Dependiendo de su ubicación será la fase de la onda reflejada.

En la figura C.3a se muestran 4 átomos contenidos dentro de una celda unitaria y los planos de Bragg más simples (con índices de Miller 0 1).

Los átomos azul y cian justo coinciden con los planos de Bragg, al ser reflejados los rayos están en fase y se suman. En cambio, el átomo rosa se encuentra exactamente entre los 2 planos de Bragg y refleja la luz desfasada $\lambda/2$ provocando interferencia destructiva con la reflexión de los planos de Bragg.



(a) Difracción provocada por un cristal complejo.



(b) Representación vectorial de las ondas de salida.

Figure C.3. Diagrama de la reflexión que provoca un cristal de proteína.

En la figura C.3b está la representación vectorial de todas las reflexiones de C.3a. La magnitud (intensidad) de cada vector depende del número de electrones que fueron excitados por la onda

incidente, o sea el tamaño del átomo. La dirección es la fase con la que son reflejadas las ondas, o sea la posición del átomo dentro de la celda unitaria.

La contribución total del conjunto de planos de Bragg resulta ser la suma de todos los vectores de reflexión. Esta contribución corresponde al complejo C_{hkl} de la ecuación 2.1.

Por último, se van a dar algunos ejemplos utilizando imágenes tomadas de la página de la universidad de York [27]. La figura C.4 da información del factor de estructura de un cristal. La ventana de arriba a la izquierda índica la reflexión seleccionada; muestra sus índices de Miller y la amplitud y fase de la onda reflejada. El cuadro de abajo muestra la representación vectorial de la onda total reflejada. En la ventana de arriba a la derecha está el espacio recíproco; el eje a^* equivale al índice de Miller h y el b^* a k. Cada punto representa la reflexión total sobre los planos de Bragg correspondientes, la intensidad índica la amplitud, y el color la fase (que no se conoce directamente). Finalmente en la ventana de abajo a la derecha, el fondo rojo y azul representa la función de densidad $\rho(x, y)$ del cristal, la caja gris claro con lados a, b es la celda unitaria y las lineas paralelas que la cruzan son los planos de Bragg seleccionados. La onda verde es la onda total reflejada.

En la figura C.5 y C.6 se tienen ejemplos de seis conjuntos de planos. Arriba de cada una se muestra una representación del patrón de difracción que genera el cristal. Está acomodado de acuerdo al espacio de los índices de Miller (h k) en dos dimensiones que es un espacio con simetría sobre el centro, es decir, el punto (hk) y el $(-h-k)$ forman a la misma familia de planos de Bragg.

En la figura C.5 se representan tres distintos conjuntos de planos de Bragg que provocan una reflexión total de la onda incidente. Se puede ver que en todos los casos la separación entre los dos átomos es igual a la distancia entre los planos de Bragg.

En la figura C.6 se tienen tres casos contrarios en los que los planos no provocan ninguna reflexión total. En todos se tiene un átomo que coincide con los planos de Bragg y otro que esta en medio, esto provoca que las dos ondas reflejadas se destruyan.

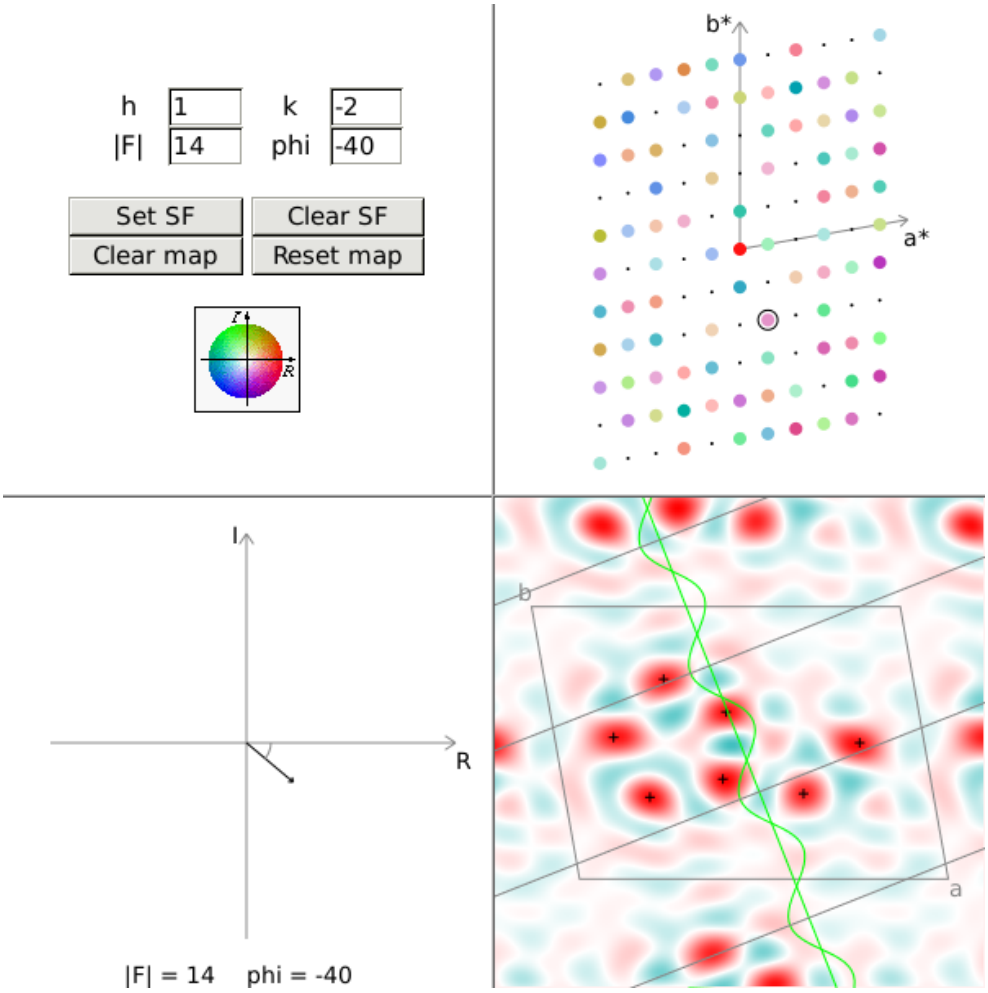
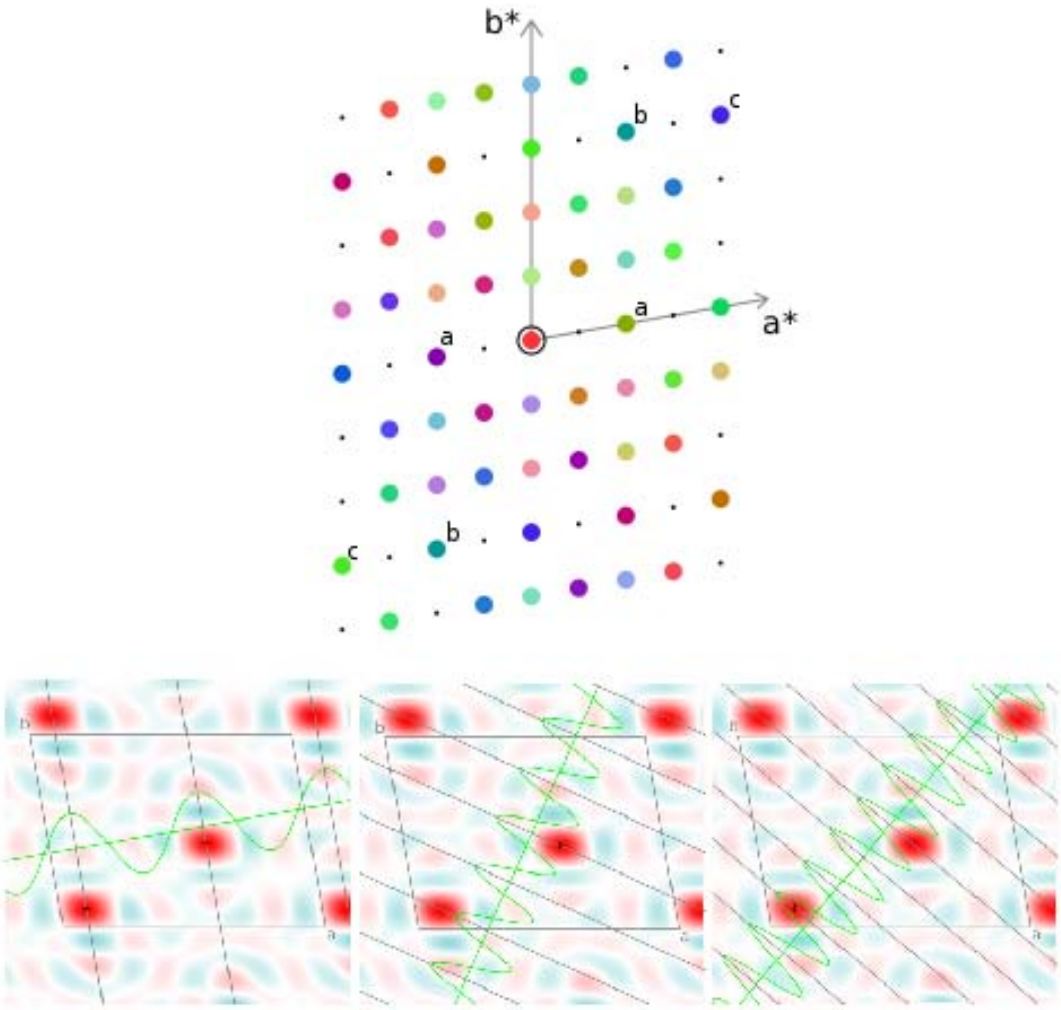


Figure C.4. Contribución para el conjunto de planos correspondiente a los índices de Miller 1 -2.

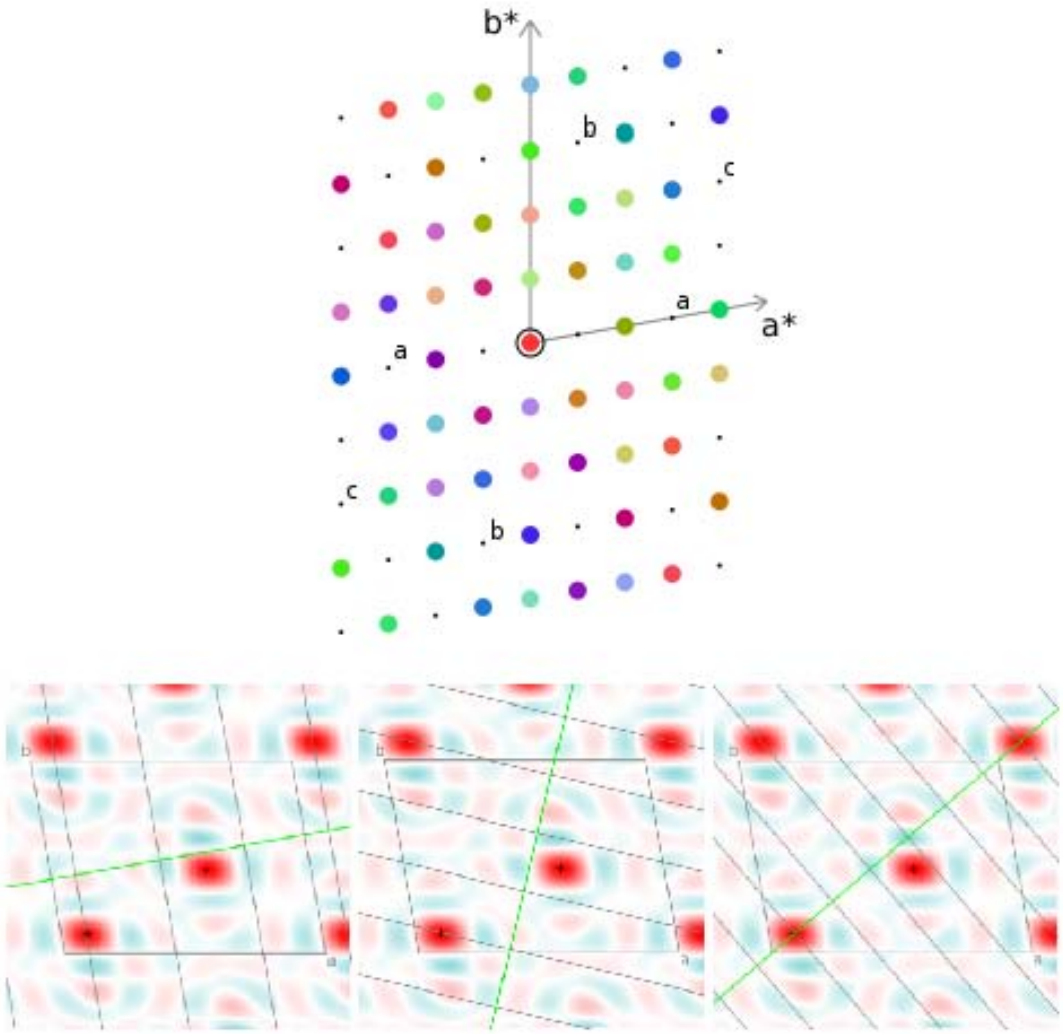


(a) h=2 k=0 con $\phi = 72$ y su recíproco.

(b) h=2 k=3 con $\phi = 180$ y su recíproco.

(c) h=4 k=3 con $\phi = -108$ y su reciprocó.

Figure C.5. Planos de Bragg que provocan una reflexión total por lo que su intensidad es equivalente al número total de electrones que hay en cada unidad |F| = 12



(a) h=3 k=0 y su recíproco. (b) h=1 k=3 y su recíproco. (c) h=4 k=2 y su reciprocó.

Figure C.6. Planos de Bragg que no provocan ninguna reflexión, en todos los casos un átomo coincide con los planos de bragg y el otro esta exactamente en medio.

Bibliografía

[1] Lehninger (2005). *Principles Of Biochemistry* (4a ed.). E.U.A NY: W.H. Freeman and Company.

[2] Pierre-Yves Calland (2003). *On the structural complexity of a protein*. Protein Engineering vol.16 no.2 pp.79–86.

[3] Dorn, Barbachan e Silva, Buriol, Lamb (Diciembre 2014). *Three-dimensional protein structure prediction: Methods and computational strategies* Computational Biology and Chemistry, Volume 53, Part B, Pages 251–276.

[4] Torres, A. M., Menz, I., Alewood, P. F., Bansal, P., Lahnstein, J., Gallagher, C. H., Kuchel, P. W. (2002). *D-Amino acid residue in the C-type natriuretic peptide from the venom of the mammal, Ornithorhynchus anatinus, the Australian platypus*. FEBS Letters. 524 (1–3): 172–6.

[5] Kreil G (Abril 1994).*Peptides containing a D-amino acid from frogs and molluscs.* The Journal of Biological Chemistry 269(15):10967-70.

[6] Marco Ambriz-Rivas, Nina Pastor, Gabriel del Rio (Marzo 2012). Biochemistry, Genetics and Molecular Biology. Editado por: Jianfeng Cai and Rongsheng E. Wang, ISBN, InTech. *Relating protein structure and function through a bijection and its implications on protein structure prediction* (Capitulo 18).

[7] Nicholas Sperelakis (2012) *Cell Physiology Sourcebook, Essentials of Membrane Biophysics* (4a ed.). E.U.A, Ohio: Elsevier.

[8] Antonio Rey, Jeffrey Skolnickt (1991). *Efficient algorithm for the reconstruction of a protein backbone from the -carbon coordinates* Journal of Computational Chemistry.

[9] Paula Yurkanis Bruice (2005). *Organic Chemistry* (4a ed.). E.U.A, Pennsylvania: Prentice Hall.

[10] Siegfried Haussühl (2007). *Physical Properties of Crystals* Germany, Cologne: Wiley-VCH.

[11] Ricardo Corral-Corral, Edgar Chavez, Gabriel Del Rio (2015) *Machine Learnable Fold Space Representation based on Residue Cluster Classes*. Computational Biology and Chemistry 59(A) 1-7.

[12] Xin-She Yang (2008). *Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics*. E.U.A, Cambridge: Cambridge International Science Publishing.

[13] David Rutten *Simulated Annealing, a brief introduction*. [en línea] (Octubre 2012) [fecha de consulta: 23 Noviembre 2016] Disponible en: https://ieatbugsforbreakfast.wordpress.com/2011/10/14/simulated-annealing-a-brief-introduction/

[14] S. Kirkpatrick; C. D. Gelatt; M. P. Vecchi (1983). *Optimization by Simulated Annealing*. Science, New Series, 220(4598). pp. 671-680.

[15] Smith, Temple F. Waterman, Michael S. (1981). *Identification of Common Molecular Subsequences.* Journal of Molecular Biology. 147: 195–197.

[16] Bernal, J. D., Crowfoot, D. *Nature 133, 794–795* 1934.

[17] J. C. Kendrew et al. (1958). *X-ray Photographs of Crystalline Pepsin*. Nature 181, 794-795.

[18] Perutz M. F. (1960) *Structure of haemoglobin* Brookhaven symposia in biology. 13: 165–83.

[19] Yang,Y. et al. (2012) *A new size-independent score for pairwise protein struc- ture alignment and its application to structure classification and nucleic-acid binding prediction.* Proteins, 80, 2080–2088.

[20] Peter Brown, Wayne Pullan, Yuedong Yang, Yaoqi Zhou (Octubre 2015). *Fast and accurate non-sequential protein structure alignment using a new asymmetric linear sum assignment heuristic* Bioinformatics 1-8.

[21] National Center for Biotechnology Information. *GenBank and WGS Statistics* [en línea] Rockville Pike (2016) [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.ncbi.nlm.nih.gov/genbank/statistics/.

[22] The Protein Data Bank. *PDB Statistics* [en línea] (2016) [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html.

[23] Kamlesh Sahu. *Classification of Amino Acid* [en línea] Kbiotech International Applied Science Network (2016) [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.kbiotech.com/Tm/aaclassi.htm.

[24] University of Washington. *Rotamer Library* [en línea] (2011). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.dynameomics.org/Rotamer/indexRotamer.aspx

[25] Drs. J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, Springer Nature. *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis* [en línea] (2006). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.nature.com/physics/looking-back/kendrew/index.html

[26] Tony Phillips, American Mathematical Society. *X-ray Crystallography and the Fourier Transform.*[en línea] (2011). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.ams.org/samplings/feature-column/fc-2011-10

[27] Dr Kevin Cowtan, University of York. *The Interactive Structure Factor Tutorial*. [en línea] (Octubre 2014). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.ysbl.york.ac.uk/ cowtan/sfapplet/sfintro.html.

[28] Bernhard Rupp. *Multiple Isomorphous Replacement (MIR) Phasing* [en línea] (Diciembre 2009). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.ruppweb.org/Xray/MIR/mirphasing.html.

[29] I. Sillitoe, N. Dawson, T. Lewis, D. Lee, J. Lees, C. Orengo, Protein Structure Classification Database. *CATH* (2016). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www.cathdb.info/.

[30] Yuzhen Ye, Adam Godzik. *Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists*. [en línea] (2015). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://fatcat.burnham.org/

[31] Andreas Prlic, Philip Bourne. *Combinatorial Extension*. [en línea] (2013). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://source.rcsb.org/jfatcatserver/ceHome.jsp.

[32] Víctor Marintell García. *Get protein 3D structure from RCC vector* [en línea] (2016). [fecha de consulta: 23 Noviembre 2016] Disponible en: https://github.com/purinlord/3dforrcc.

[33] Randy J Read, University of Cambridge. *Protein Crystallography Course* [en línea] (Junio 2005). [fecha de consulta: 23 Noviembre 2016] Disponible en: http://www-structmed.cimr.cam.ac.uk/course.html.

[34] Dr. Kevin Ahern (2016). *Alpha helix structures*, *Beta strands*, *Antiparallel beta*. [Figura]. Recuperado de: http://oregonstate.edu/instruct/bb450/fall13/lecture/proteinstructureIoutline.html.

[35] *X-Ray diffraction photograph of a single crystal of sperm whale myoglobin.*. [Figura]. Recuperado de: https://www3.nd.edu/ aseriann/CHAP8B.html/sld027.htm.

[36] (2011) *unión peptidica* [Figura]. Recuperado de: http://franmorancyp.blogspot.mx/.

[37] Westfälische Wilhelms, Universität Münster (2012) *aminosäuren liste* [Figura]. Recuperado de: http://d-nb.info/1029742855/34.

[38] Biology Questions and Answers (2016). *Protein-Synthesis* [Figura]. Recuperado de: http://www.biology-questions-and-answers.com/protein-synthesis.html.

[39] StructureFunction *Specify the four structural levels of proteins...* [Figura]. Recuperado de: http://vickypang.weebly.com/unit-2.html.