**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

Maestría en Ciencias Bioquímicas

# Evolución *in silico* de interacciones proteína ligando: profilina/poli-L-prolina como modelo de estudio

## TESIS

**QUE PARA OPTAR EL GRADO DE:**

**Maestro en Ciencias Bioquímicas**

## PRESENTA:

Jorge Enrique Quintana Kageyama

**DR. ENRIQUE MERINO PÉREZ**

**INSTITUTO DE BIOTECNOLOGIA**

**CUERNAVACA,MORELOS NOVIEMBRE 2010**

# Índice

## *Agradecimientos*

La presente tesis de maestría se enmarca dentro del proyecto de colabaroción de los grupos de investigación de los doctores Federico Sánchez y Enrique Merino, ambos del Instituto de Biotecnología-UNAM; por tal razón deseo agradecer de manera particular al Dr. Federico Sanchez y al Dr. Gabriel Guillén, del laboratorio del Dr. Sánchez ya que sin ellos no hubiera sido posible realizar mi proyecto de maestría.

Tambien quisiera agradecer al Dr. Enrique Rudiño por su ayuda en el análisis de las estructuras tridimensionales, las discusiones tenidas con él y su apoyo.

Finalmente, a mi tutor el Dr. Enrique Merino Pérez, que ha sido un gran guía durante este tiempo, y en particular la gran paciencia y apoyo que me ha dado.

## *Introducción*

Todos los organismos vivos tienen la característica de responder a una gran cantidad estímulos endógenos y ambientales. Estas respuestas tienen que ser rápidas y específicas para cada grupo de estímulos; priorizar el metabolismo de ciertas fuentes de carbono; la expresión temporal de genes de desarrollo; respuestas a agentes tóxicoss, entre otras. Este tipo de respuestas son cruciales para la supervivencia del organismo.

Para poder generar respuestas de esta naturaleza, el organismo necesita alguna vía de comunicación entre la célula y sus alrededores, es decir, se necesita alguna forma de sensar el ambiente, amplificar la señal y finalmente generar la respuesta deseada. Uno de los mecanismos principales que tienen los organismos para dicha tarea es la interacción y formación de complejos de proteínas con otras moléculas.

Las proteínas son cadenas lineales de aminoácidos; estas cadenas pueden ser divididas dominios funcionales, que llevan a cabo funciones específicas. Son estos dominios los que llevan a cabo las interacciones proteína-ligando. Sin embargo, no todos los aminoácidos dentro de estos dominios son necesarios para la correcta union de sus ligandos; solo algunos de ellos determinan la fuerza y específicidad de dicha interacción [1].

Identificar los residuos que llevan a cabo estas interacciones es de gran interés debido a su potencial uso terapéutico e industrial, ya que con ello se pueden predecir compuestos químicos que podrían inhibir la formación de estos complejos e inhibir sus mecanismos de acción.

Actualmente, mediante la técnica computacional llamada acoplamiento molecular, es posible obtener predecciones de las estructuras de los complejos proteína-ligando y obtener funciones de energía a partir de las estructuras tridimensionales de cada una de las moléculas que interactúan entre si. Hoy en día, incluso existen bibliotecas de estructuras de compuestos químicos pequeños que se usan para hacer búsquedas de posibles inhibidores de proteínas.

Pero ésto solo es el comienzo, el poder entender los mecanismos que permiten la correcta interacción entre una proteína y su ligando tiene como consecuencia que

potencialmente se pueda usar este conocimiento para hacer modificaciones a una proteína, generando variantes que tengan mayor afinidad por su ligando, o inclusive una especificidad alterada por éste. A este campo de estudio se le conoce como "Diseño de Proteínas".

El diseño de proteínas es un problema muy desafiante debido a la gran cantidad de variables que se tienen que considerar, el gran espacio de búsqueda y la complejidad en el comportamiento de las proteínas. Sin embargo, podemos encontrar en la literatura múltiples casos en el que se ha tenido éxito, desde crear proteínas *de novo* hasta la modificación de proteínas existentes mejorando ciertas propiedades. Por citar algunos ejemplos de lo anterior, el grupo del Dr. David Baker, en la universidad de Washington, se logró diseñar una proteína *de novo* con capacidad de realizar una reacción Diels-Adler (reacción para generar anillos de 6 carbonos usando un dieno y una olefina simple ) reacción que ninguna proteína conocida en la naturaleza la puede realizar [2].Ademas, el grupo del Dr. Steve Mayo del Instituto de Tecnología de California, ha generado un programa capaz de predecir cambios que estabilizan proteínas [3]; el grupo de la Dra. Birte Höecker del Instituto Max-Planck de Biología de Desarrollo, logró darle actividad catalítica a una proteína Barril-$(\beta\alpha)_8$ artificial mediante Diseño Racional; un método para el diseño de proteinas que consiste en la mutación de aminoácidos cercanos al sitio de interacción [4].

Como éstos, hay un gran número de casos exitosos en el área de diseño de proteínas, cada uno usando estrategias diferentes, que van desde un diseño puramente computacional, hasta diseños racionales únicamente observando la estructura tridimensional de las proteínas.

Nuestro proyecto de investigación considera que es posible realizar la evaluación computacional sistemática de mutaciones en una proteína para que, con base en valores de afinidad predichos mediante acoplamiento molecular para un ligando específico, sean seleccionadas aquellas variantes con una afinidad mayor respecto a la proteína silvestre. Es decir, nuestro proyecto tiene como objetivo realizar la *evolución in silico* de una proteína para aumentar su afinidad por su ligando natural.

## *Antecedentes*

La evolución *in silico* de proteínas es un tema de reciente creación dentro del área de Ingeniería de Proteínas. Es hasta hace poco que ha sido empleada con éxito logrando aumentar las afinidades de diversas proteínas por sus ligandos, cambiando su especificad, o bien aumentando su estabilidad. Los aspectos teórico-computacionales involucrados en la evolución *in silico* de proteínas son fundamentalmente los siguientes:

### *Predicción de estructuras tridimensionales de proteínas*

La predicción de la estructura tridimensional de proteínas es un tema que ha sido sumamente estudiado [5] y se ha dividido en 3 categorías:

I) Análisis comparativo, en donde se intenta obtener la estructura de una proteína al compararla con otra proteína relacionada de la cual se tiene su estructura tridimensional resuelta; a esta última se le llama templado.

II) *Threading* o "hilado", en donde la predicción consiste en alinear a cada uno de los aminoácidos en la secuencia problema con una posición determinada en la proteína templado cuya estructura es conocida, y se evalúa qué tan bien se ajusta dicho modelo de acuerdo a las propiedades estadísticas de las proteínas depositadas en la Base de Datos de Proteínas (PDB). La predicción mediante hilado molecular puede efectuarse aun cuando no exista ninguna relación evolutiva entre la proteína problema y las proteínas templado.

III) Predicción *de novo*, en donde los modelos predicen la estructura de la proteína sin ninguna información *a priori* adicional a la secuencia primaria de la proteína.

Existe un gran desarrollo de algoritmos computacionales en estas 3 categorías, siendo los algoritmos con mayor éxito aquellos que usan templados en sus predicciones.

### *Acoplamiento molecular*

Hace 30 años Wodak y Janin [6] desarrollaron el primer sistema de acoplamiento molecular. Desde entonces el problema de acoplamiento molecular ha sido un tema muy estudiado y se han generado algoritmos de búsqueda con enfoques muy diversos, como pueden ser: i) correlaciones de las transformadas de Fourier, ii) *hashing* geométrico y iii) modelos de Monte Carlo. Sin embargo a pesar de las mejoras, estos algoritmos generan muchas soluciones para la orientación del ligando, entre las cuales se encuentran múltiples falsos positivos. Estos algoritmos en general tienen problemas identificando el mecanismo real de estas interacciones ante todas las opciones que encuentra.

En adición al problema anteriormente mencionado, el acoplamiento molecular generalmente asume una estructura rígida de la proteína para poder simplificar la gran cantidad de cálculos que se tienen que hacer. El método generado por Katchalski-Katzir [7] ha ayudado a resolver el problema de poder de cómputo, generando un algoritmo que puede ser ejecutado eficientemente en una computadora moderna. Inicialmente, este algoritmo solo estaba basado en la complementariedad de formas entre el sustrato y el ligando, posteriormente dicho algoritmo se ha mejorado tomando en cuenta diferentes parámetros como son las interacciones electrostáticas que existen entre ambas moléculas.

Actualmente hay muchos algoritmos diferentes para resolver el problema del acoplamiento molecular, cada uno usando métodos de muestreo diferentes, tomando en cuenta la flexibilidad de la proteína, y diferentes formas de calificar la fuerza de las interacciones entre el ligando y la proteína.

### *Diseño computacional de proteínas*

El diseño computacional de proteínas es un área de estudio de mucho interés. La sola idea de poder generar proteínas *de novo* que tengan una actividad catalítica, es sin duda una idea prometedora. Generar una proteína *de novo* con la función deseada ha sido un problema que no ha tenido solución universal, hay pocos casos que han logrado generar estas proteínas con éxito, y siempre han necesitado de rondas extras de

evolución dirigida. Sin embargo, el mejoramiento o cambio de la función de una proteína que ya se existe en la naturaleza ha sido un campo que ha tenido mayor éxito [8].

Uno de los factores que más ha influido en el diseño computacional de proteínas ha sido el gran desarrollo de la inteligencia artificial de los últimos años [9]. El problema de diseño de proteínas radica en que se tienen que explorar espacios de búsqueda muy grandes, es decir, hay una cantidad exorbitante de posibles variantes que existen sobre la estructura de una proteína, sin embargo solo nos interesa aquellas que cumplen nuestros criterios funcionales, que son una gran minoría. El desarrollo de la inteligencia artificial es una herramienta que nos ayuda a explorar estos espacios de manera mucho más inteligente, en un tiempo razonable.

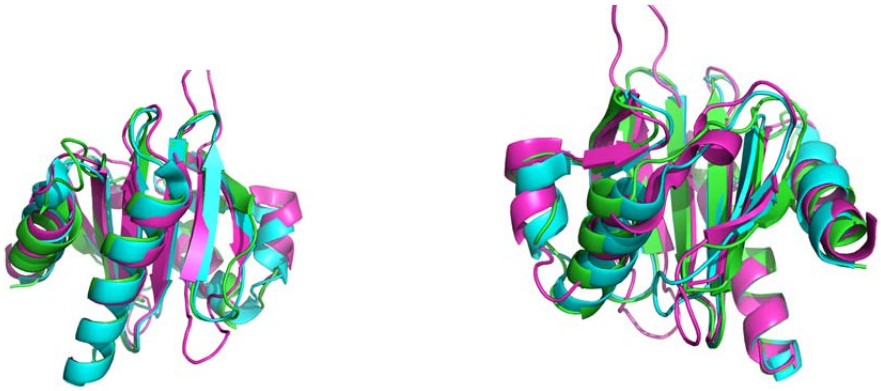
Uno de los mejores ejemplos de la eficiencia del uso de inteligencia artificial aplicado al diseño de proteínas es el caso de RosettaDesign [8] y Egad [10]. Rosetta usa algoritmos de Montecarlo para reducir los choques hidrofóbicos de la proteína, esperando que como consecuencia mejore su función, mientras que Egad usa un algoritmo genético para muestrear mutaciones que mejoren la estabilidad de la proteína.

### *Profilina*

Como se mencionó anteriormente, la predicción de la estructura de proteínas y la modificación de función de una proteína es un problema más sencillo si se tiene la estructura de ésta o al menos un templado de una proteína homóloga o estructuralmente relacionada (se infiere esta relación si hay un porcentaje alto de identidad a nivel de secuencia) . Debido a estas razones y por su gran relevancia biológica en organismos eucariontes, en nuestro proyecto hemos escogido a profilina como proteína de estudio. Profilina es una proteína que regula la formación del citoesqueleto de actina en la célula. Esto lo hace secuestrando monómeros de actina. Sin embargo, profilina también favorece el intercambio de ADP-ATP en la actina, aumentando su nivel de polimerización. Aparte de estos mecanismos moleculares, profilina tiene mayor afinidad por monómeros de actina-ATP que por monómeros de actina-ADP, generando un fino mecanismo de regulación [11]. Profilina es una proteína

con roles centrales de regulación en todos los organismos eucariontes, lo cual hace a ésta una proteína de gran interés biológico.

A.



**B.**

| % de identidad | Arabidopsis | Humano | Levadura |
|---|---|---|---|
| Arabidopsis | %100 | %24.1 | %30.3 |
| Humano | %24.1 | %100 | %26.2 |
| Levadura | %30.3 | %26.2 | %100 |

**C.**

| RMSD | Arabidopsis | Humano | Levadura |
|---|---|---|---|
| Arabidopsis | 0 | 1.91 | 1.52 |
| Humano | 1.91 | 0 | 1.95 |
| Levadura | 1.521 | 1.95 | 0 |

**Figura 1. Estructuras de profilinas sobrepuestas de tres especies distintas. En verde, rosa y azul encontramos las estructuras de profilina de Arabidopsis, humano y ratón respectivamente.** Se puede ver que a pesar de no tener un porcentaje de identidad alto son estructuralmente muy parecidas.

Profilina tiene características muy particulares que la hacen una excelente molécula modelo para proyectos de investigación dentro del área de ingeniería de proteínas:

- La familia de profilinas se encuentran en todos los eucariontes, y tiene estructuras tridimensionales muy similares entre si, a pesar de que no estén tan conservadas a nivel de secuencia (Figura 1.B). Lo anterior puede apreciarse al sobreimponer las diferentes estructuras obtenidas por cristalografía de rayos X.(Figura 1.A) Estas características son de particular valor ya que es factible poder hacer buenas

predicciones de la estructura de una profilina modificada al hacer modelaje comparativo.

- Uno de los ligandos de profilina es poli-L-prolina (PLP), este es un péptido pequeño compuesto de 14 residuos de prolina. La formación del complejo proteína ligando ha sido caracterizada en diferentes variantes de profilina, donde también se han medido las afinidades por dicho sustrato. Estos resultados permiten evaluar la capacidad predictiva de resultados teóricos ya que podemos contrastar nuestras predicciones contra los resultados experimentales
- Experimentos de mutación sitio-dirigida han demostrado que mutaciones en ciertos sitios pueden aumentar o disminuir al afinidad de profilina por sus ligandos bilógicos.

Entre las variantes de profilina mejor caracterizadas se encuentra profilina de *Arabidopsis thaliana* [12]; hay varias estructuras cristalográficas(PDB 1A0K,3NUL) y muchos estudios sobre el mecanismo de acción de ésta, por lo que la hemos escogido como modelo inicial para nuestro estudio.

Considerando lo anteriormente expuesto, nuestro proyecto de investigación plantea el uso de diferentes algoritmos de ingeniería de proteínas para identificar aquellas mutaciones puntuales que favorezcan la afinidad de profilina por su ligando PLP, considerando las siguientes hipótesis y objetivos:

## *Hipótesis*

1) Tomando como molde la estructura tridimensional de una proteína, se pueden predecir con gran precisión las estructuras tridimensionales de sus variantes aún que difieran en pocos aminoácidos.

2) Se puede predecir computacionalmente la estructura del complejo proteína-ligando y evaluar la afinidad de dicho complejo mediante funciones de energía.

3) Se pueden utilizar algoritmos computacionales de búsqueda para obtener una combinación de mutaciones puntuales sobre una proteína para que, en teoría, presenten una mayor afinidad por su ligando.

# Objetivos

## Objetivo General

Desarrollar una metodología computacional automatizable usando acoplamiento molecular e inteligencia artificial para seleccionar aquellos cambios puntuales en una proteína que aumenten la afinidad por su ligando.

## Objetivo particulares

1) Evaluar las capacidades y eficiencia de los algoritmos actuales que se emplean para la predicción de estructura, y la predicción de interacciones proteína-ligando.

2) Utilizar algoritmos de inteligencia artificial para implementar una metodología computacional automatizada que permita predecir cambios puntuales en las proteínas que aumenten la afinidad por sus ligandos naturales.

3) Aplicar la metodología desarrollada para encontrar variantes de profilina de *Arabidopsis thaliana*, que aumenten su afinidad por PLP.

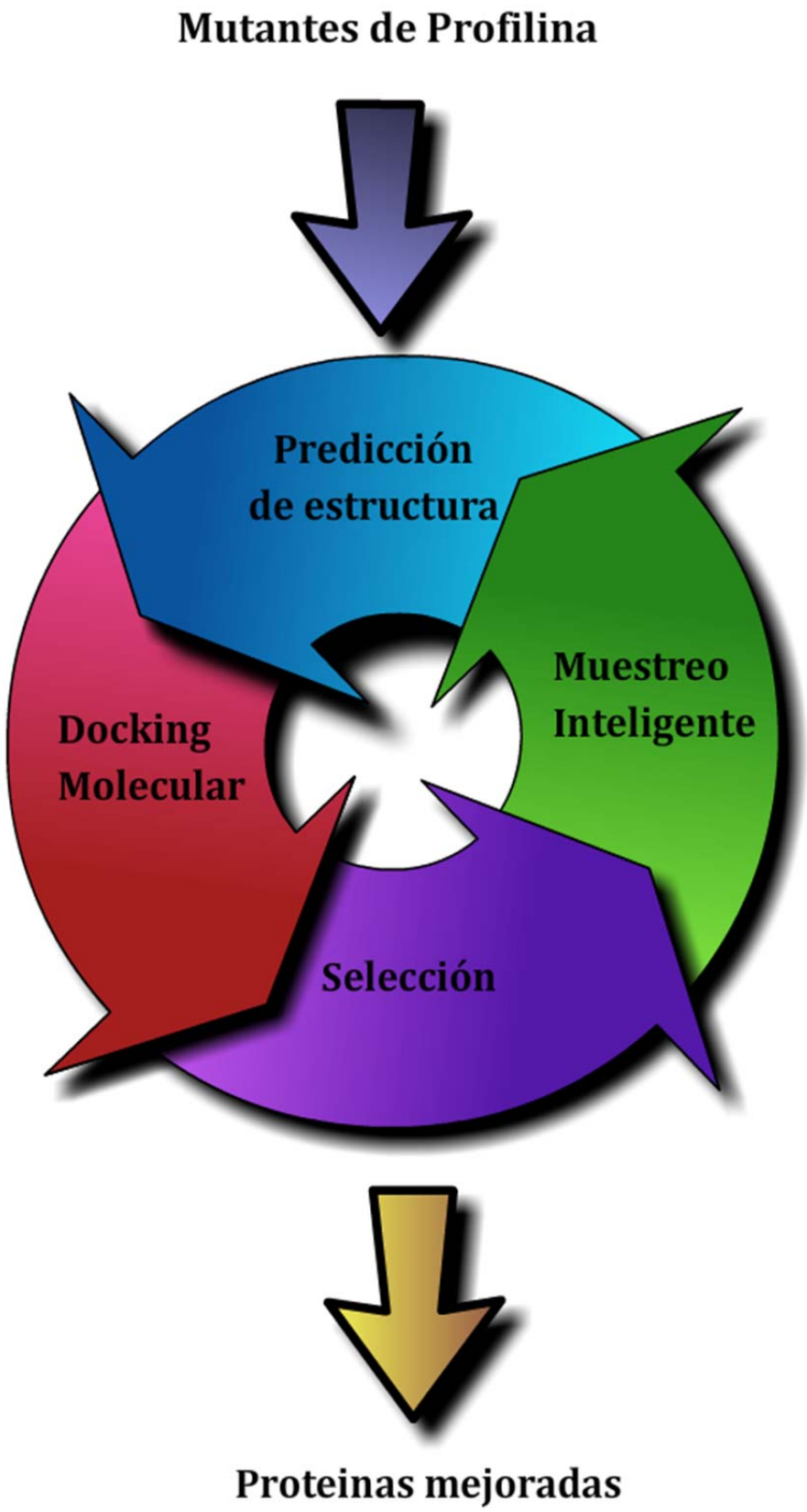
## *Metodología*

Primero se evaluarán la calidad de las predicciones de estructura mediante diferentes programas, entre ellos *Modeller* [13] y *Rosetta* [8]. Se necesitará analizar la calidad de estas predicciones para evaluar si son lo suficientemente confiables para generar una predicción del complejo ligando-proteína. Para esta tarea, se predecirá la estructura de diferentes variantes de profilina de las cuales ya se cuente con estructura tridimensional usando únicamente la secuencia de aminoácidos de estas, y se comparará con la estructura obtenida experimentalmente.

Se usarán diferentes programas de cómputo para realizar el análisis de acoplamiento molecular y se evaluarán los resultados obtenidos con resultados ya probados experimentalmente, así observando que tan confiables son estos resultados. Ejemplos de este tipo de programa son, *RosettaDock* [8], *GRAMM* [7].

Se generarán algoritmos de inteligencia artificial para poder muestrear grandes espacios de búsqueda dadas nuestras limitaciones en poder de cómputo. Hay una gran cantidad de posibles combinaciones de mutaciones de una proteína, sin embargo, el efecto que tiene una mutación sobre la estructura y afinidad, es un efecto cooperativo y no-aditivo, ya que las interacciones no son independientes. Por esta razón se pueden generar algoritmos para buscar máximos locales sin tener que explorar todo el espacio de secuencias que existen. Ejemplos de estos se encuentran algoritmos de enjambres [14], modelos de Montecarlo [15] y algoritmos genéticos [16].

Una vez evaluada las capacidades de estos programas, se seguirá la siguiente metodología para generar variantes de profilina ilustrada en el siguiente diagrama;




1. **Muestreo Simple**: Se iniciará con la generación de variantes de profilina generadas por mutaciones puntuales al azar con el uso de código en Perl.

2. **Predicción de estructura**: Se predecirán las estructuras de estas variantes usando modelaje comparativo.

3. **Acoplamiento molecular**: Se generarán predicciones de las afinidades de estas proteínas por su respectivo ligando mediante el uso de acoplamiento molecular.

4. **Selección:** Se usarán criterios de selección para elegir las proteínas que hayan mejorado la afinidad por su ligando *in silico*.

5. **Muestreo Inteligente**: Se usarán algoritmos de inteligencia artificial para generar nuevas variantes tomando como referencia las proteínas escogidas en el paso 4. Las variantes nuevas pasarán por los pasos 2,3,4 y 5 hasta que no encontremos una variante que posea una mayor afinidad predicha

### *Muestreo Simple*

Para decidir el tamaño del primer muestreo, primero tenemos que conocer el tamaño de nuestro problema. Profilina tipo II de *Arabidopsis thaliana* tiene 131 residuos. Tomando en cuenta que hay 20 variantes de aminoácidos, la cantidad de permutaciones de una sola mutación es muy sencilla de calcular: 131 residuos x 20 aminoácidos=2,620 combinaciones posibles.

Usando una computadora portátil con un procesador Intel Dual Core a 1.8 MHz, el tiempo de cómputo para la predicción de una estructura por homología en promedio es de 10 segundos; el tiempo requerido para un análisis promedio de acoplamiento molecular es de 5 segundos, dándonos un total de 15 segundos por variante, por lo que para analizar el conjunto de mutaciones sencillas en una computadora estándar se requeriría:

$$2620 \text{ variantes x } 15 \text{ segs} = 39{,}300 \text{ segs} = 9{,}825 \text{ min} = 163 \text{ hrs} \sim 7 \text{ días}$$

Este es un tiempo muy razonable, en particular porque es un problema altamente paralelizable, es decir se puede correr las predicciones en varias computadoras al mismo tiempo.

Cuando consideramos el caso de mutaciones dobles, el tiempo requerido para realizar las permutaciones de 20 aminoácidos por residuo es:

$$\binom{131}{2} \times 20^2 = \frac{131!}{2! \times (131-2)!} \times 20^2 = 3{,}406{,}000\ combinaciones \times 15\ segs$$
$$= 51{,}090{,}000 segs \sim 591\ días$$

Como podemos ver la cantidad de combinaciones es increíblemente grande, inclusive para sólo dos mutaciones. Por esta razón queda descartado el análisis del total de dobles mutaciones de nuestro problema.

Otra de las ventajas que tenemos usando profilina como modelos de estudio es que se conocen los dominios responsables para el pegado a PLP en la proteína[8]. En particular se ha demostrado que las dos estructuras de alfa hélice en el amino y carboxilo terminal son suficientes para el pegado de PLP [17]. Si sólo usáramos estas regiones, nuestro espacio de búsqueda se reduce drásticamente:

$$\binom{29}{2} \times 20^2 = \frac{29!}{2! \times (29-2)!} \times 20^2 = 162400 \times 15\ segs = 1436000 segs \sim 26.5\ dias$$

A pesar de que esto resulta en un tiempo considerable, puede ser reducido aun más si se excluyen ciertos aminoácidos del análisis y si se usan múltiples computadoras.

### *Predicción de estructura*

Para esta tarea se escogió el programa de *Modeller 9v8* [13]. El algoritmo del programa Modeller fue descrito por Andrej Sali en 1993. Este algoritmo se basa en satisfacer restricciones espaciales. La información de entrada es un alineamiento y la estructura templado, y el resultado es un modelo tridimensional de la proteína de interés con posicionamiento de cadenas principales y laterales. Este algoritmo se basa en la siguiente estrategia(Figura 2).

1) Alineamiento: Se genera un alineamiento entre la proteína de la cual se quiere predecir su modelo y aquella proteína de la cual ya se conoce su estructura tridimensional. Es importante notar que el tipo de alineamiento generado es un alineamiento especial ya que considera información estructural de la proteína de la que se tiene su estructura, esto ayuda a forzar residuos de mayor importancia estructural a alinearse.

2) Extracción de restricciones espaciales:
   a) Una vez generado este alineamiento se obtienen restricciones espaciales de una biblioteca que fue generada previamente haciendo análisis estadísticos de las bases de datos de estructuras tridimensionales. Estas restricciones espaciales

están expresadas como funciones condicionales de densidad de probabilidad o pdf's. En si son las probabilidades de ver objetos a cierta distancia dada la naturaleza de los objetos.

b) Después se calculan múltiples términos (longitud de enlace, ángulo de enlace, ángulos dihedrales impropios basados en los campos de fuerza CHARM-22) [18], intentando forzar así una estereoquímica correcta, finalmente estos términos de energía son combinados con las restricciones espaciales generadas en el paso dos, generando una función objetivo.

3) El modelo final es generado usando funciones de gradiente y recocido simulado para la optimización de dicha función objetivo.

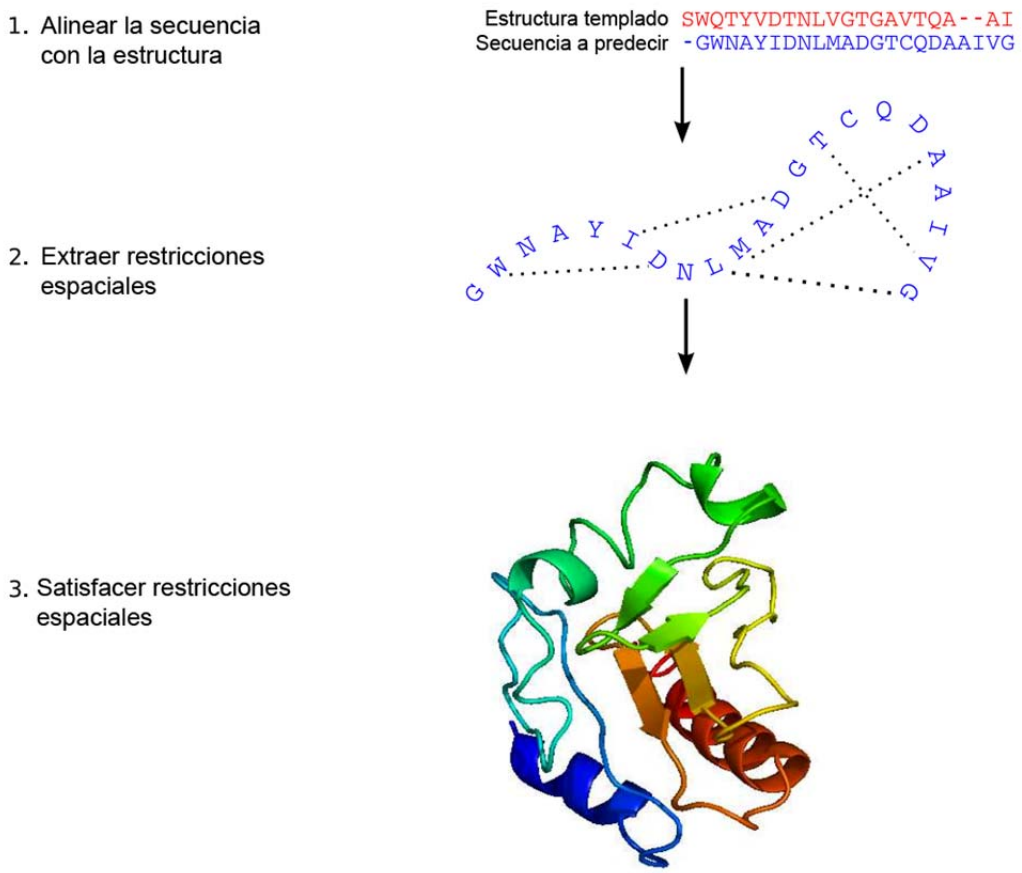



**Figura 2. Flujo de trabajo que sigue el Programa Modeller 9v8**. Primero se alinean las secuencias a un templado. Se obtienen restricciones espaciales entre los aminoácidos que componen la proteína y finalmente genera una estructura que intenta satisfacer la mayoría de estas restricciones espaciales.

Una vez generados nuestros modelos tridimensionales podemos hacer un paso de refinado adicional para mejorar la calidad de éste. Para esta tarea se decidió utilizar el programa *Crystalography and NMR System* o *CNSolve*[11]. Este programa fue originalmente diseñado para aplicaciones en problemas de Resonancia Magnética Nuclear y Cristalografía de rayos X, sin embargo también tiene herramientas para optimizar estructuras tridimensionales.

Esta minimización se lleva a cabo en dos pasos:

1. El primer paso de la optimización consiste en verificar que no haya ningún choque entre las cadenas de carbonos alfa; en caso de haberlo, modifica la estructura para eliminarlos.

2. En el segundo paso, fija los carbonos alfa en sus coordenadas tridimensionales e intenta optimizar el posicionamiento de las cadenas laterales de la proteína mediante un algoritmo de recocido simulado.74

El algoritmo de recocido simulado (SA por sus siglas en inglés: *Simulated Annealing*) es un algoritmo meta-heurístico para la búsqueda de óptimos globales en espacios discretos muy grandes. Este está inspirado en el templado de materiales en la industria metalúrgica mediante el calentamiento y enfriado controlado de materiales para la obtención de cristales más grandes y reducción de impurezas. En sí, la excitación de átomos a temperaturas altas les permite explorar una gran cantidad de estados energéticos diferentes, el enfriado controlado les permite encontrar otros estados con menor energía.

Análogamente en el SA se tiene un parámetro K que se quiere optimizar. En cada paso del algoritmo se escoge una nueva solución que depende de dos cosas, la diferencia de una función de energía entre los dos estados y de un parámetro global llamado temperatura. Entre más alta sea la temperatura, el parámetro a optimizar aceptara con mayor facilidad un cambio. Cuando el parámetro de temperatura empieza a decrecer,

entonces las soluciones dejan de ser menos aleatorias y empiezan a dejar de variar. Una temperatura alta permite escapar de mínimos locales, mientras que una temperatura baja, explora soluciones cercanas(Figura 3).

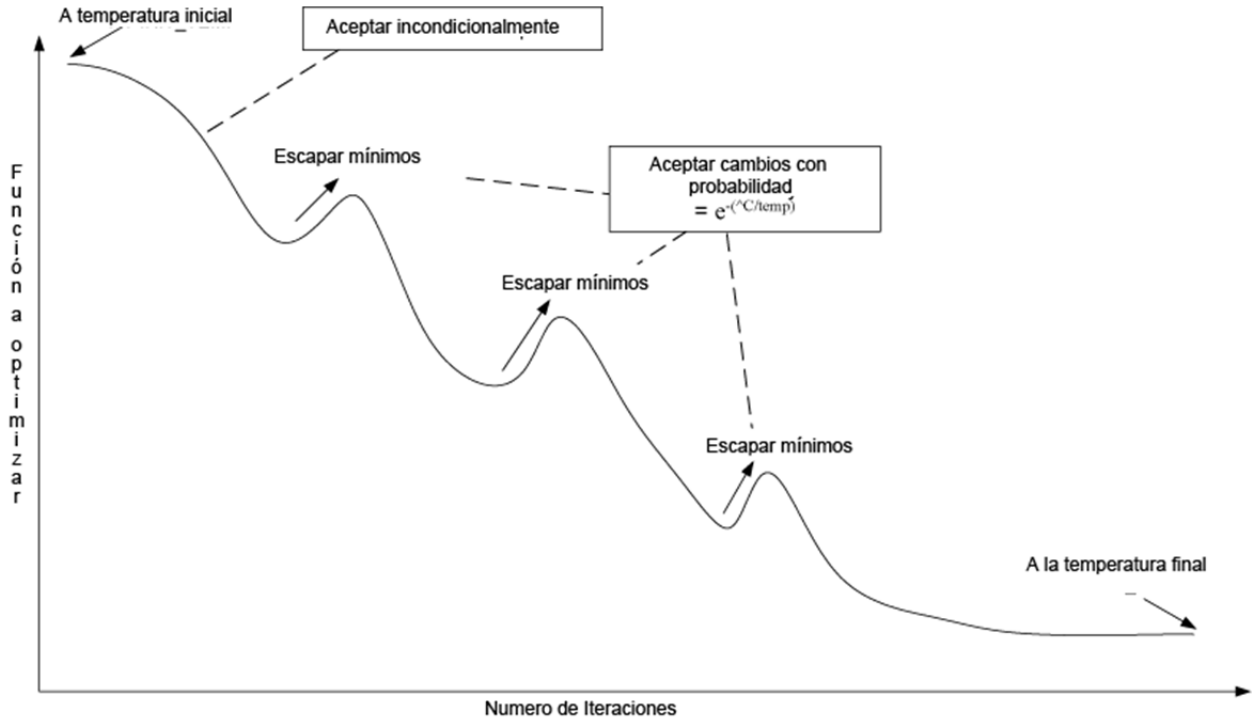



**Figura 3. Visualización del algoritmo de recocido simulado.** Se puede buscar los óptimos globales sin quedarse "atrapado" en óptimos locales [19].

### *Acoplamiento molecular*

Para hacer un acoplamiento molecular hay una gran variedad de programas disponibles. Este tipo de programas tienen una gran gama de estrategias para poder identificar el sitio y forma de pegado entre un ligando y su proteína. Algunos algoritmos buscan maximizar la superficie de contacto, mientras que otros buscan encontrar un mínimo energético en la proteína. Debido a este hecho, cada estrategia tiene sus

ventajas y sus desventajas, sin mencionar que varía mucho el tiempo de cómputo que cada uno requiere.

Debido a que nuestro ligando es una cadena polipeptídica que consiste en una hélice de 14 prolinas, es probable que la fuerza motora de la interacción con profilina sea más geométrica, que electroestática. Debido a esto, nuestra primera opción para generar modelos de acoplamiento entre PLP y profilina fue el uso de un programa de acoplamiento geométrico llamado GRAMM.

GRAMM [7] es un programa de reconocimiento de superficies para la identificación de complementariedad de superficies entre una proteína y un ligando. Este algoritmo usa técnicas de reconocimiento de patrones y rápidas transformadas de Fourier. El algoritmo consiste en contraponer las dos estructuras en un cubo tridimensional, y evaluar cada punto del cubo con una función de correlación.

Primero definimos la función ā y b:

$$\bar{a}_{l,m,n} = \begin{cases} 1 & \text{en la superficie de la molécula} \\ \rho & \text{adentro de la molécula} \\ 0 & \text{afuera de la molécula} \end{cases}$$

$$\bar{b}_{l,m,n} = \begin{cases} 1 & \text{en la superficie de la molécula} \\ \delta & \text{adentro de la molécula} \\ 0 & \text{afuera de la molécula} \end{cases}$$

Donde l, m, n son las coordenadas dentro del cubo. La función de correlación que se quiere optimizar es la siguiente:

Ecuación 1

$$\bar{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \bar{a}_{l,m,n} \cdot \bar{b}_{l+\alpha,m+\beta,n+\gamma},$$

Donde m, n y γ son cambios de coordenadas.

Como podemos ver, se evalúan todos los puntos dentro del cubo. Si hay múltiple coordenadas positivas, la multiplicación de estas también va a ser positiva, por lo que la función de correlación aumenta. Sin embargo, solo queremos que la interacción entre las superficies aumente la función de la correlación, si existe una penetración entre el ligando y la proteína buscamos que esto disminuya la función de correlación. Para lograr esto, asignamos valores negativos muy grandes a **ρ**, y valores positivos muy pequeños a δ en las funciones a y b. De esta manera, si b penetra la estructura de a, la multiplicación de **ρ**, por 1 o δ, va a ser un número negativo, reduciendo el valor de correlación.

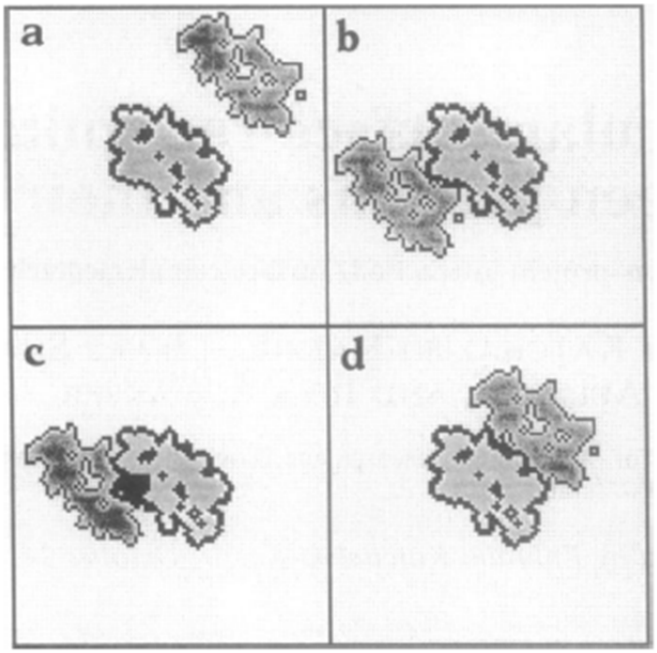



**Figura 4. Ejemplos de posiciones relativas de a y b.**
(a) No hay contacto, c=0;
(b) Contacto limitado c pequeño

(c)Penetración c=-k

(d)Una buena complementariedad geométrico.

En la ecuación 1 vemos que las únicas coordenadas que son cambiadas para hacer el cálculo de la función de correlación son las coordenadas del ligando. Sin embargo, no solo podemos incluir translaciones sobre el ligando, también hay que considerar las rotaciones, esto nos genera un espacio $N^6$ dimensional. Para reducir el tiempo de cómputo, se usaron las transformaciones rápidas de Fourier para el cálculo de la función de correlación, reduciéndolo de $N^6$ a $N^3 \log N^3$.

El resultado de GRAMM es la estructura de tridimensional con el mejor acoplamiento encontrado, y un número que representa el grado de complementariedad entre el ligando y la proteína.

## Resultados

### Predicción de estructura

Como primer paso, se evaluó la capacidad del programa *Modeller* para realizar la predicción de estructuras de profilina. Para esta tarea se escogieron 3 profilinas de diferentes organismos de las cuales ya se tienen resueltas sus estructuras cristalográficas: *Arabidopsis thaliana*, *Homo sapiens* y *Sacharomices cerevisae* (PDB: 1A0K,1D1J,1K0K respectivamente).

Para cada proteína con estructura, se extrajo la secuencia de aminoácidos de esta, y se uso para generar una predicción de estructura de la misma, usando como templado la estructura de profilina de otro organismo. Esta predicción se uso para compararla con la estructura cristalográfica, para determinar que tan similar es una predicción con la estructura real. Por ejemplo, se predijo la estructura de profilina de humano usando su secuencia de aminoácidos y como templado la estructura de profilina de levadura. Esta predicción fue después comparada con la estructura determinada por cristalografía de humano.

El RMSD o *Root Mean Square Deviation* es la raíz cuadrada de la media de la diferencia de posiciones entre carbonos alfa equivalentes de dos estructuras. Si todos los

carbonos alfa de dos estructuras están en la misma posición, el RMSD de estas tiene un valor de cero.

Generalmente es considerado que la predicción de una estructura con un RMSD menor a 3 **Å** con respecto a la estructura real es una buena predicción. En nuestro resultados obtenemos RMSD's muchos menores a este valor (ver Tabla 1). Debido al hecho de que el modelaje comparativo mejora entre más cercanas sea la proteína de interés y la proteína de la cual se tiene un modelo, podemos asumir que los modelos de nuestro estudio con tan sólo dos mutaciones van a ser mucho mejores.

**RMSD entre profilinas.**

| | **Arabidopsis** | **Humano** | **Levadura** |
|---|---|---|---|
| **Arabidopsis** | 0.133 | 1.675 | 1.46 |
| **Humano** | 2.065 | 0.273 | 1.899 |
| **Levadura** | 1.502 | 1.862 | 0.135 |

**Tabla 1. RMSD en Å entre diferentes modelos tridimensionales de profilinas**. Cada estructura de las diferentes 3 especies fue modelada usando como templado cada una de las estructuras cristalográficas depositadas en PDB. Estos modelos fueron comparados con las posiciones de los carbono alfa de lo modelos depositados en PDB. Los valores de las diferencias se dan en unidades RMSD.

Para probar la calidad de nuestro programa de acoplamiento molecular se predijeron los complejos de PLP y profilina de ratón del cual ya se tiene estructura cristalográfica. Se obtuvo una predicción con un RSMD de 0.97 con respecto a la proteína de ratón. El programa también colocó correctamente PLP en el dominio de pegado de profilina(Figura 6).

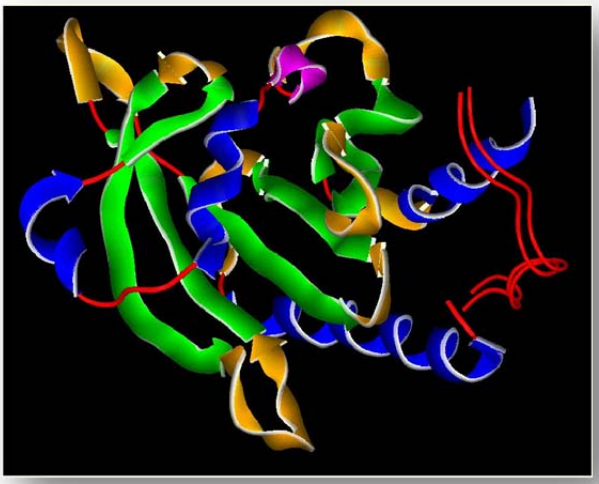


**Figura 6. Modelamiento de la unión de profilina de Rata con el complejo PLP**. La estructura del complejo PLP y Profilina de Rata fue separado en dos estructuras, y usando acoplamiento molecular se recuperó una estructura sumamente similar a la original (RMSD 0.97)

### *Predicción de Complejo proteína-ligando*

Habiendo probado que tenemos una buena calidad en nuestros modelos de estructura y en la predicción de complejos, se generaron todas las variantes sencillas de profilina tipo I de *Arabidopsis thaliana* y se modelaron tomando como molde a la estructura tridimensional de profilina tipo I de *Arabidopsis thaliana* (pdb ID 1A0K).

Una vez predichos dichos modelos, se generó una segunda biblioteca con los mismos modelos después de una segunda etapa de refinación con CNSolve 1.3[11]. Se usaron ambas bibliotecas como entradas para el programa de acoplamiento molecular para poder comparar el impacto de la minimización sobre los resultados del acoplamiento. Una vez teniendo los resultados de GRAMM, se generaron tablas con todos los resultados y se graficaron como mapas de calor(Figura 7).

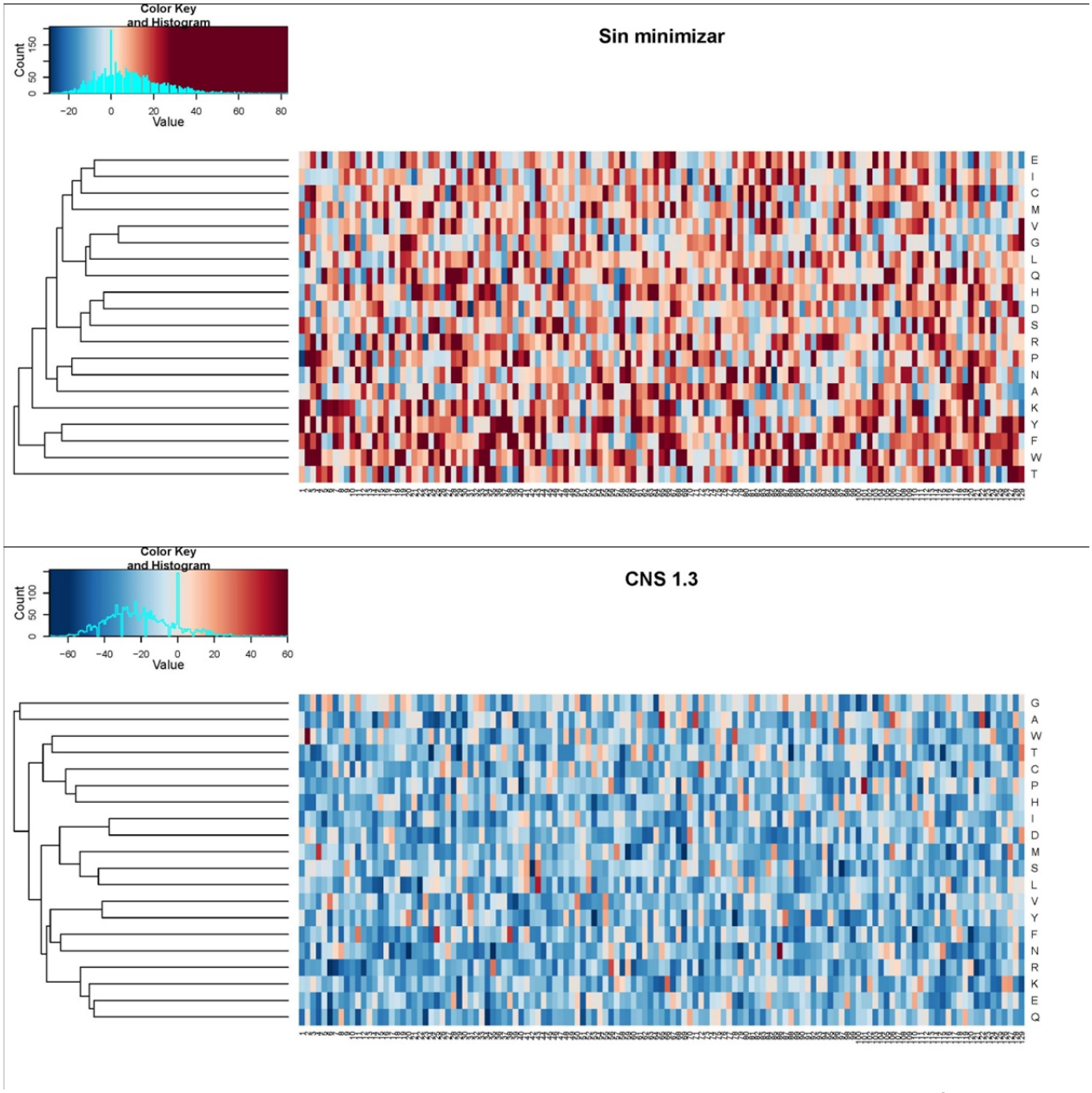



**Figura 7. Mapas de calor de los valores de acoplamiento de las variantes de profilina a PLP como ligando**. Se graficaron los valores de GRAMM(unidades arbitrarias de complementariedad) de todas las variantes de un solo cambio de aminoácido y se le restó el valor de la proteína nativa. Azul corresponde a un valor menor a la proteína nativa, mientras que rojo indica un mejor ajuste que la proteína nativa. El eje de las X representa la posición del residuo.

Como podemos ver en la Figura 7, hay una gran variabilidad entre los resultados de GRAMM dentro de cada conjunto. Cuando se comparan los mejores valores de las predicciones de complementariedad entre las dos bibliotecas, las predicciones no son consistentes(Figura 8). Los anteriores resultados nos da un indicio del efecto que tiene el refinamiento sobre las predicciones en el acoplamiento.

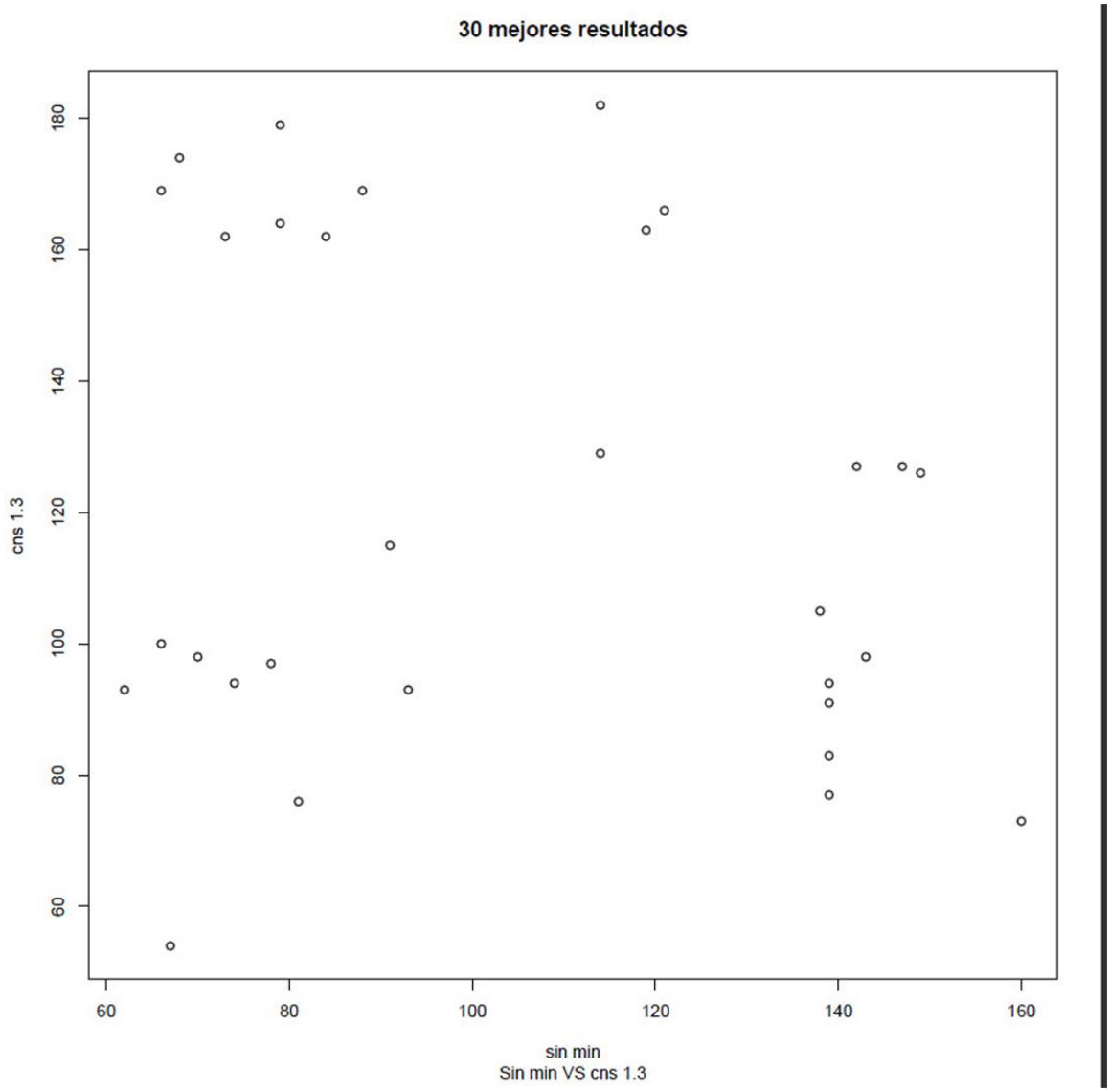



**Figura 8. Comparación de resultados de acoplamiento de modelos de profilina con y sin refinamiento a PLP como ligando.** Se compararon los resultados de GRAMM (usa unidades arbitrarias de complementariedad) de los mejores resultados de la biblioteca(resultados con mayor complementariedad geométrica predicha por GRAMM) de modelos sin refinar, y modelos refinados.

En el caso de la biblioteca con minimizaciones, el número de mutaciones predichas que mejoran el puntaje de complementariedad geométrica es muy bajo comparado con los resultados obtenidos con la biblioteca de estructuras que no fueron refinados. Esto va en contra de resultados reportados, ya que se sabe que la mayoría de las mutaciones en una proteína son deletereas. Esta observación nos sugiere que el paso de refinamiento es esencial para poder hacer este tipo de estudios.

En estudios previos se ha demostrado que hay dos estructuras alfa hélices localizadas en el carboxilo y amino terminal de profilina respectivamente, que son suficiente para hacer el pegado a PLP. Debido a que los residuos contenidos en estas dos alfa hélices son aquellos que interaccionan en el pegado de profilina a PLP, se espera que mutaciones en estos residuos tengan un efecto en la interacción de profilina con PLP.

Al observar la varianza de los resultados podemos notar que ésta aumenta dentro de los dominios en el carboxilo terminal y el amino terminal(Figura 9), lo cual está en acuerdo con la evidencia experimental y es consistente con la idea de que cambios en estos extremos van a tener un efecto mayor en la unión del profilina con PLP. Sin embargo, también notamos que hay residuos fuera de estos dominios con varianzas altas.

(a) (b)

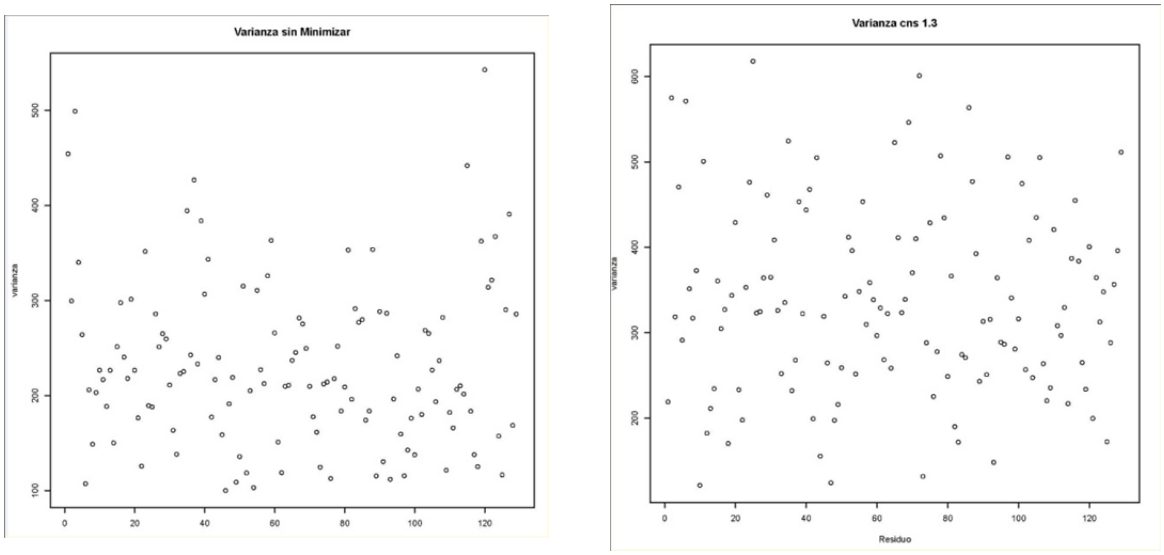



**Figura 9. Varianzas de los valores de acoplamiento de variantes de profilina a su ligando PLP.** Se graficaron las varianzas por cada sitio de la proteína. El eje de las X representa la posición del residuo, el eje de las Y es la varianza. (a) Biblioteca sin refinamiento (b) Biblioteca con refinamiento

No se ha resuelto la estructura tridimensional de ninguna profilina en plantas acoplada con PLP. Esto dificulta el análisis y la evaluación de las soluciones de nuestro método. Debido a ello, se seleccionó una segunda variante de profilina cuya estructura tridimensional acoplada a un ligando, muy parecido a PLP, es conocida.

Profilina tipo IIA de ratón tiene como ligando un dominio rico en prolina dentro de la proteína VASP (Vasodilator-stimulated phosphoprotein). Actualmente se cuenta con la estructura de profilina acoplada a este dominio (PDB 2V8C). Aunado a esto, estudios previos mediante mutación sitio-dirigida han mapeado dos sitios funcionalmente importantes para la union al dominio de VASP. Estos datos experimentales hacen a esta variante un excelente candidato para evaluar nuestra metodología.

Se repitió el mismo análisis con esta variante de profilina. Una vez obtenidos los resultados, las soluciones del acoplamiento molecular fueron superpuestas a la estructura nativa de la proteína acoplada a su ligando. Generamos mutaciones puntuales con el objetivo de mejorar la región de contacto entre el ligando y la proteína, esperando que el mecanismo de pegado sea el mismo.

Al graficar la varianza y el promedio de los puntajes obtenidos usando GRAMM por posición en la proteína(Figura 9), podemos observar que estos estimadores puntuales son muy diferentes entre posiciones.

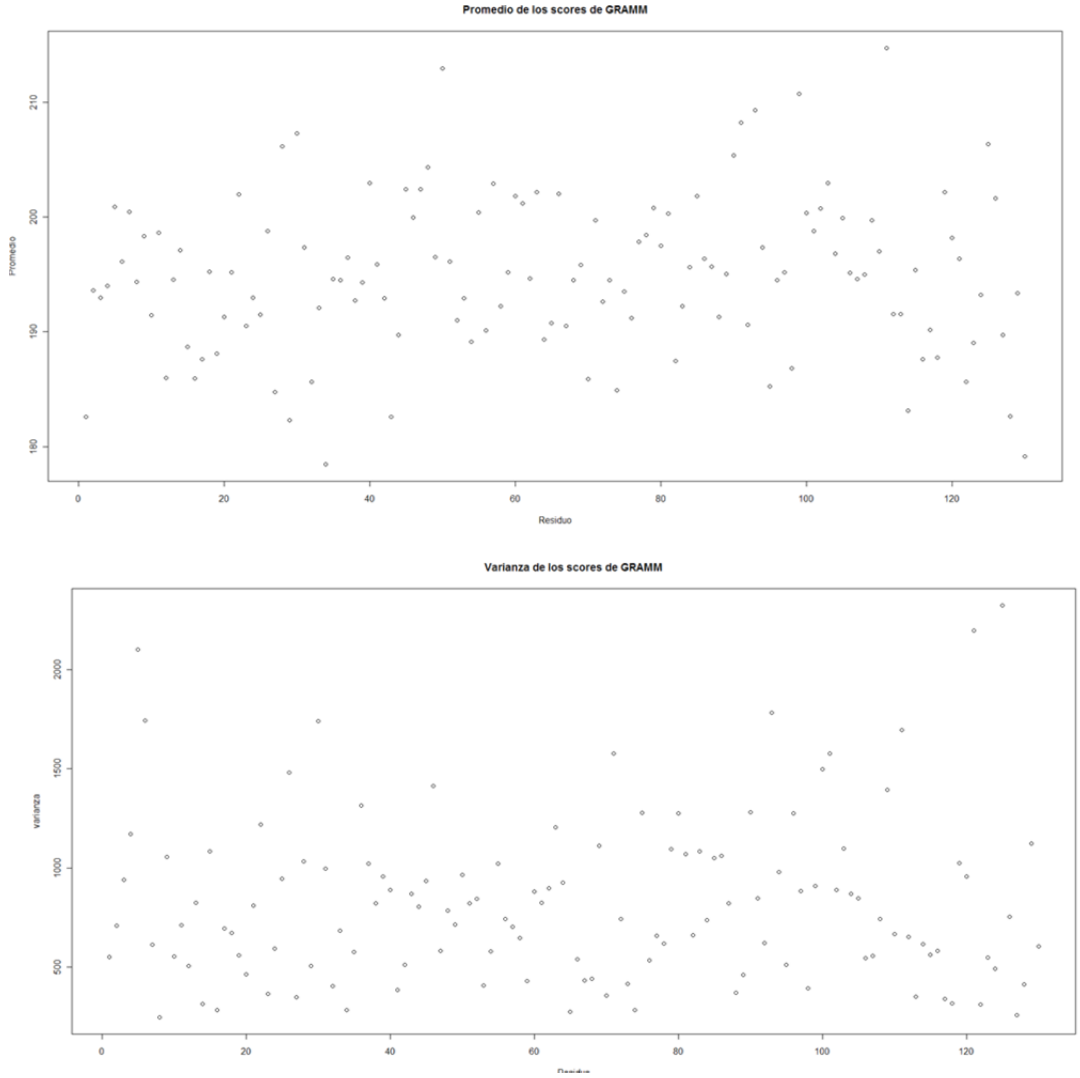



**Figura 9.** Promedio y varianza de los resultados de GRAMM por posiciones. En el eje X se encuentra la posición del residuo, y en el eje Y el promedio y varianza de los puntajes obtenidos con GRAMM

Para explicar la variabilidad de nuestros resultados de los valores de acoplamiento de nuestros modelos con su ligando, propusimos tres hipótesis:

La primera explicación contempla que alguna de las mutaciones genere un cambio en la estructura de la proteína, como el cambio de la posición de un conector generando un cambio considerable en el mecanismo de pegado, y por lo tanto en el valor reportado.

Se realizó la superposición de las soluciones con mejores y peores resultados predichos con GRAMM con respecto a la estructura de la proteína original(Figura 10), con el objetivo de observar si existía algún cambio en la estructura secundaria. Esta comparación permitió observar la diferencia de la calidad y viabilidad entre diferentes predicciones.

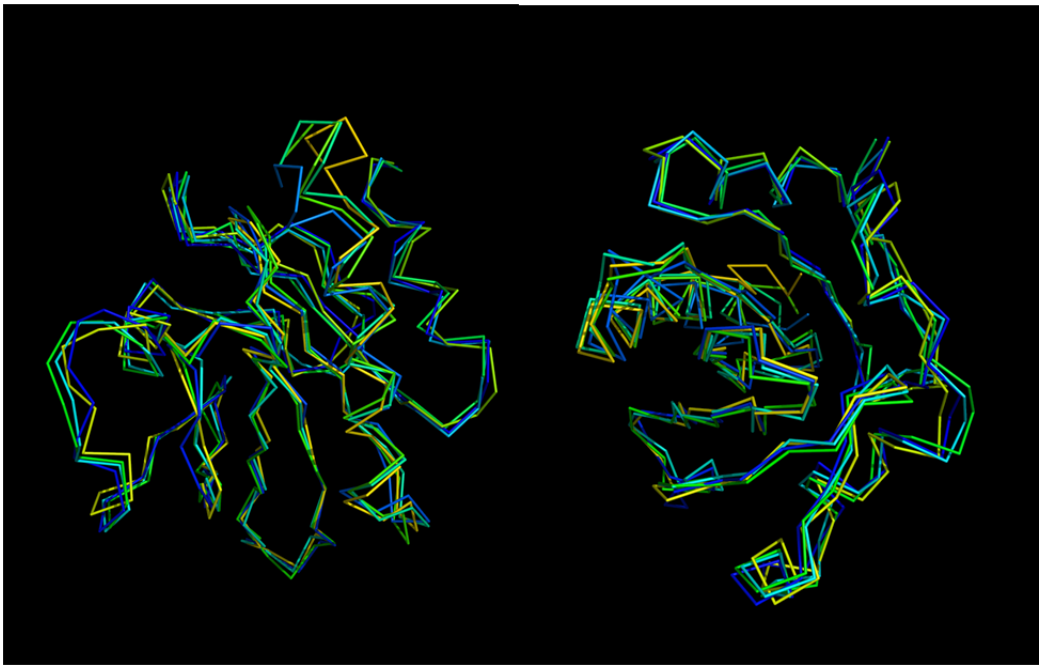
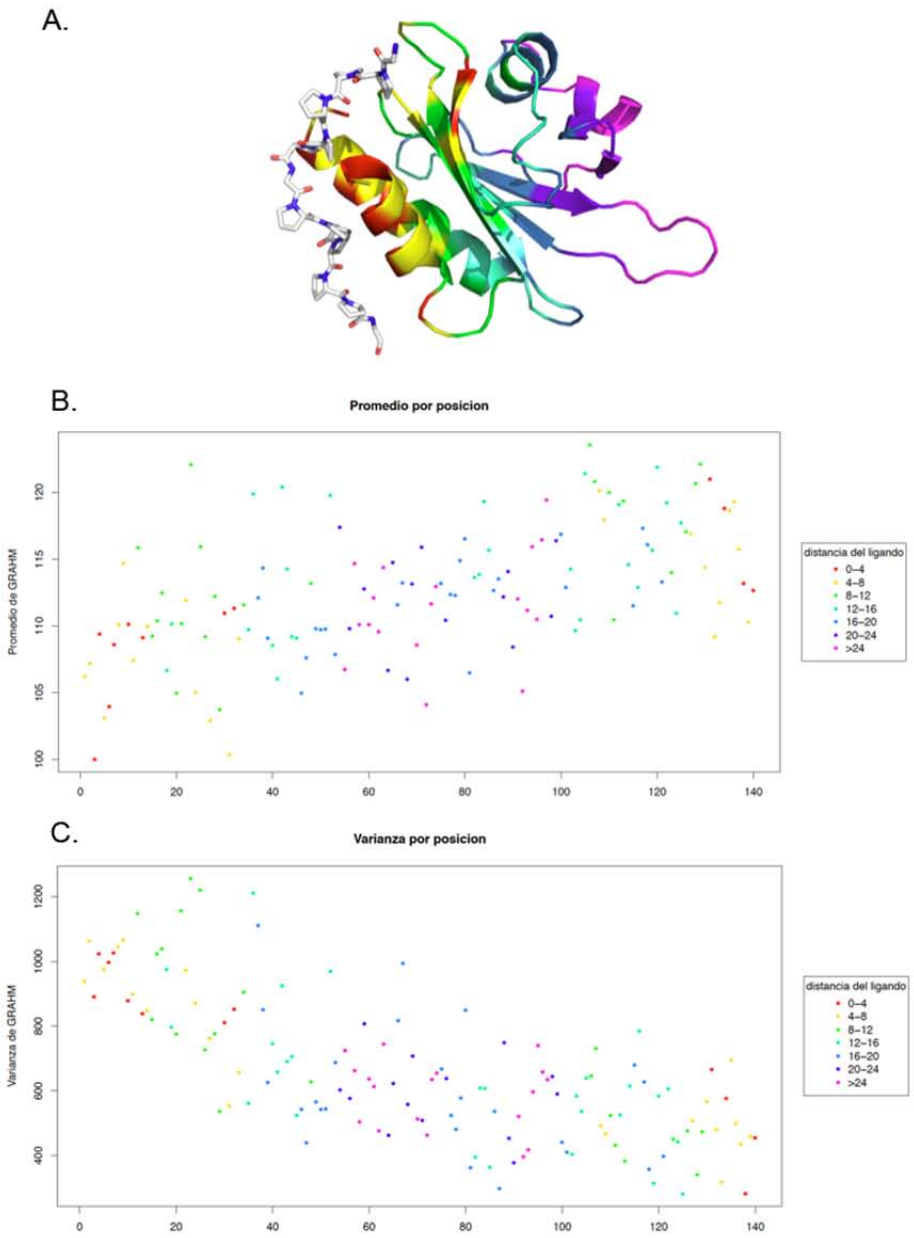


**Figura 10. Superposición de las estructuras predichas de profilina que obtuvieron los 2 mejores y 2 peores resultados de GRAMM.** Se muestran únicamente los carbonos alfa para poder apreciar cambios en el posicionamiento de estructuras secundarias.

Para explorar esta posibilidad, los diez sitios con mayor variabilidad fueron escogidos, y el RMSD de cada una de las mutantes con respecto a la estructura cristalográfica de la proteína fue calculado. En ningún caso hubo un RMSD alto (RMSD>1), lo cual indica que los carbonos alfa están en posiciones muy cercanas. Este

era un resultado esperado debido al método de predicción de estructura del programa MODELLER que genera un alineamiento a nivel de secuencia que es usado como guía. Debido a que las secuencias a modelar son 99% idénticas, el esqueleto de las proteínas




es muy similar.

**Figura 11. Promedio y varianza del score reportado de GRAMM por posicion en profilina separado por distancias al sitio de interacción.** A.Los residuos fueron clasificados con respecto a la distancia al átomo mas cercano de PLP. B. Promedio por residuo del valor reportado

por GRAMM. C. Varianza por residuo del valor reportado por GRAMM. Los colores utilizados en ls gráficas corresponden a los colores mostrados en el modelo tridimensional de la proteína.

La segunda posibilidad para explicar la variabilidad de nuestros datos de acoplamiento considera que a pesar de que las mutaciones no estén en el dominio de pegado a ligando, los sitios de mayor variabilidad están cercanos al ligando y que de esta forma se pueda generar algún tipo de impedimento estérico o cambio en la interfaz con el ligando.

Para evaluar esta hipótesis, se agruparon todos los aminoácidos en grupos de acuerdo a su cercanía con el ligando. El primer grupo consiste de todos los aminoácidos que tienen al menos un átomo a menos de 4 Å de distancia, el segundo entre 4 Å y 8 Å, el tercero 8 Å y 12 Å (Figura 11), sucesivamente hasta que todos los residuos pertenecieran a un grupo.

Si la distancia pudiera explicar la variabilidad de nuestros datos, se esperaría que cada grupo tuviera una distribución única dentro de sus resultados (varianza y media). Para probar esta idea se realizaron pruebas *Z* de dos muestras, teniendo como hipótesis nula que la distribución de resultados es independiente de la distancia de la mutación al ligando, y como hipótesis alternativa que la distribución de resultados es dependiente de la distancia al sitio de unión.

Como podemos observar en la Tabla 2, los valores p (la probabilidad de ver un valor igual o mas extremo dado que la hipótesis nula es cierta) obtenidos no son lo suficientemente significativos como para rechazar la hipótesis nula, es decir los resultados de GRAMM son independientes de la distancia. Esto es un resultado inesperado ya que suponíamos que cambios en residuos que interactúen con el ligando deberían de tener un impacto muy fuerte en el score, mientras que cambios muy lejanos, no deberían tener cambios significativos.

| Distancia en Angstroms(Å) | 0-4 | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 | >24 |
|---|---|---|---|---|---|---|---|
| 0-4 | 1.000 | 0.862 | 0.047 | 0.156 | 0.871 | 0.874 | 0.831 |
| 4-8 | 0.862 | 1.000 | 0.013 | 0.061 | 0.689 | 0.722 | 0.957 |
| 8-12 | 0.047 | 0.013 | 1.000 | 0.456 | 0.029 | 0.067 | 0.017 |
| 12-16 | 0.156 | 0.061 | 0.456 | 1.000 | 0.127 | 0.210 | 0.071 |
| 16-20 | 0.871 | 0.689 | 0.029 | 0.127 | 1.000 | 0.986 | 0.668 |
| 20-24 | 0.874 | 0.722 | 0.067 | 0.210 | 0.986 | 1.000 | 0.700 |
| >24 | 0.831 | 0.957 | 0.017 | 0.071 | 0.668 | 0.700 | 1.000 |

**Tabla 2 Valores P de una prueba zeta de dos muestras: H0=La distancia y el score son independientes H1=La distancia y el score son dependientes.** Se dividieron todas las mutaciones en grupos dependiendo de la distancia más cercana a PLP. Se obtuvo el promedio y la varianza y se hicieron pruebas z de dos muestras para cada grupo teniendo como hipótesis nula que su media y varianza son iguales(es decir los datos tienen la misma distribución). Los números son los valores p de cada prueba.

La tercera y última de nuestras hipótesis consistió en que la diferencia entre los resultados pudiera ser explicada únicamente con el posicionamiento de las cadenas laterales en la interface de la proteína con su ligando. Para analizar esta propuesta, se realizaron alineamientos estructurales de las estructuras predichas de las proteínas que tenían los 2 mejores resultados, los 2 peores resultados, la predicción de la estructura de la proteína silvestre, y la estructura cristalográfica de profilina. En la Figura 12 se encuentran dichas superposiciones. Las mutaciones con los mejores resultados se encuentran muy lejanas al sitio de pegado de profilina. A pesar de que hay numerosos casos donde se ha demostrado que una mutación lejana puede afectar el posicionamiento de las cadenas laterales en el sitio de pegado, es poco probable que este efecto esté ocurriendo múltiples veces en profilina.

En la Figura 12 también se puede observar que las soluciones que tienen los peores resultados tienen el ligando posicionado en sitios lejanos a la posición de pegado de la proteína cristalizada. Este efecto puede ser debido al posicionamiento de las

cadenas laterales en la interfaz de pegado de cada solución, y parámetros y restricciones impuestas por el programa de acoplamiento molecular.

**Figura 12. Superposición tridimensional de los modelos de las estructuras correspondientes a las dos peores y dos mejores soluciones de acoplamiento.** En gris se encuentra la posición del ligando del cristal. En turquesa y morado los mejores resultados, en azul y amarillo los peores resultados. La posición de la mutación está del mismo color dentro de la proteína

Se modificaron los parámetros de GRAMM para hacer una búsqueda más fina y con mayores restricciones. Entre los cambios más importantes se aumentó la resolución de nuestra búsqueda (eta: 2.0→1.7 Å), se aumentó la penalización de choques entre átomos (ro: 6.0→30), y debido a que ya contábamos con la orientación correcta de

nuestro ligando, se redujeron los ángulos de rotación del ligando (ai: ang12.dat→ang10.dat). Podemos ver en la Figura 13 que las soluciones mejoran drásticamente, ya no tenemos choques entre cadenas y las soluciones están mejor posicionadas con respecto a la ubicación esperada de pegado.

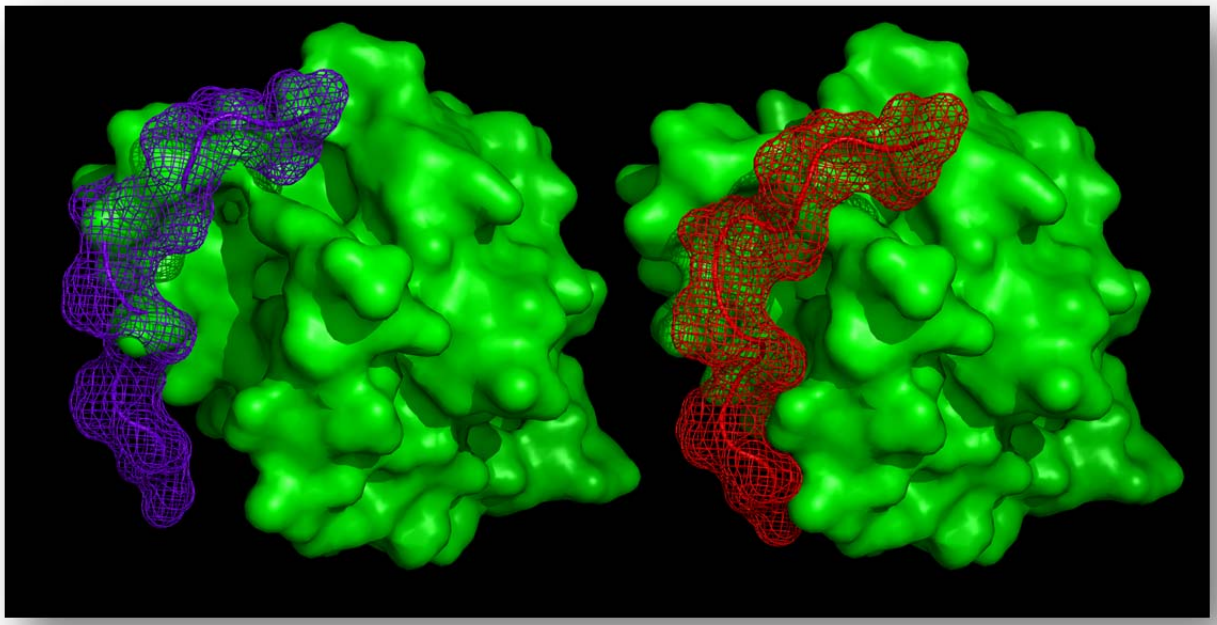


**Figura 13. La posición predicha de pegado por GRAMM del ligando con la proteína nativa**. En morado usando los parámetros predeterminados. En rojo después de modificaciones a los parámetros de búsqueda.

Usando estos mismos parámetros optimizados, se analizaron nuevamente todas las mutaciones sencillas dentro de la proteína. Al observar en la Figura 14 las posiciones predichas del ligando con resultados más altos, podemos ver que el programa está posicionando el ligando en lugares diferentes al sitio de pegado de profilina a pesar de que este posicionando correctamente el ligando en la proteína silvestre. Debido a esto necesitamos forzar las soluciones a estar en la vecindad del sitio conocido de pegado de profilina.

**Figura 14. El complejo profilina/PLP predicha por GRAMM de las 10 estructuras con resultados más altos**. En verde del lado derecho se encuentra la posición del PLP del cristal.

Desafortunadamente GRAMM no tiene opciones para hacer restricciones espaciales de posiciones del ligando, es decir no podemos forzar al programa a posicionar el ligando en lugares cercanos al sitio de pegado de profilina. Sin embargo, lo que podemos hacer es seleccionar únicamente los residuos que sean parte del sitio de unión y usarlos para los cálculos del acoplamiento molecular. Para esto se escogieron todos los residuos que tuvieran al menos un átomo a una distancia de 8 **Å** de cualquier átomo del ligando en la estructura de cristal, y todos los demás residuos fueron eliminados.

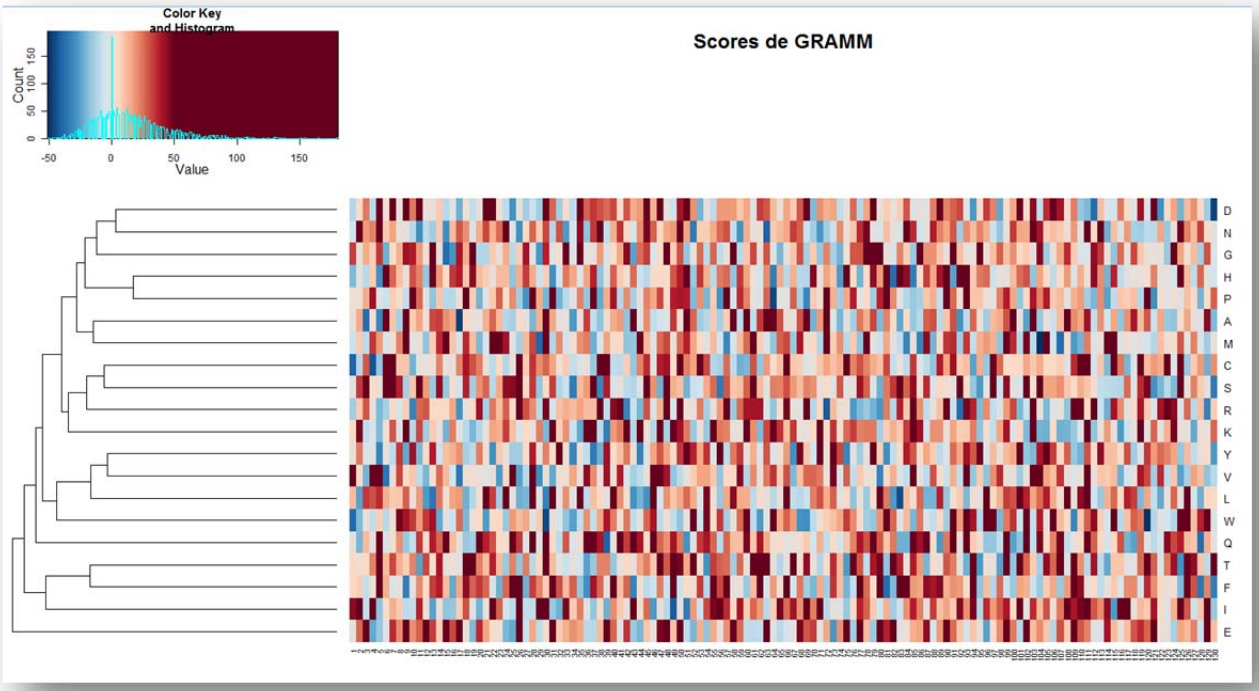



**Figura 15. Mapa de calor con los resultados de las mutaciones sencillas en profilina únicamente utilizando los residuos pertenecientes al sitio de pegado**. En rojo y en azul están las mutaciones que recibieron un score de Acoplamiento molecular más alto y más bajo respectivamente.

## *Discusión*

Al analizar los resultados obtenidos con GRAMM con nuestros nuevos parámetros y estructuras, podemos ver que la mayoría de las mutaciones obtuvieron un score más alto que los residuos de la proteína silvestre(Figura 15). Esta predicción va en contra de la evidencia experimental demostrando que la gran mayoría de las mutaciones en una proteína son deletéreas.

Hay una observación muy importante que hacer, pero para esto necesitamos definir las siguientes notaciones:

1.- A la estructura silvestre, llamémosla Wt.

2.-Ahora, tomemos cualquier mutante y llamémosla X. Vamos a denotar como P(X) a todos los residuos pertenecientes a X que se encuentran dentro de una distancia (8 Å) del ligando, en nuestro caso PLP.

3.- Tomemos el caso de dos variantes de profilina generadas por mutaciones sencillas que no estén dentro del sitio de pegado natural de su ligando, es decir que su mutación esté a mayor distancia que 8 Å de PLP. Designemos a dichas variantes A y B, respectivamente.

4.-Por definición, las mutaciones de las variantes A y B no se pueden encontrar en P(A) y P(B) (Las mutaciones tienen que estar a mas de 8 Å , pero P(A) y P(B) son los residuos que están a menos de 8 Å.)

5.- Como Wt no tiene mutaciones (punto 1), y las mutaciones de A y B no se encuentran en P(A) ni en P(B), es fácil ver que P(Wt), P(A) y P(B) tienen exactamente los mismos residuos.

6.-Hay que recordar que en nuestros últimos resultados, para cada variante (X), solo se usaron los residuos con una distancia menor a 8 Å del ligando (P(X)).

Volvamos a las variantes A y B (punto 3), cuando comparamos los resultados A y B, realmente estamos comparando los resultados P(A) y P(B) (punto 6), pero como ya habíamos dicho antes P(A), P(B) y P(Wt) tienen exactamente los mismos aminoácidos (punto 4 y punto 5).

Esto es una observación muy importante, porque implica que al comparar los resultados de GRAMM de aquellas estructuras que tengan mutaciones a mas de 8 Å de distancia del ligando, estamos comparando estructuras de proteínas que son idénticas a nivel de aminoácidos. Previamente ya habíamos notado que la posición de los carbonos alfa están conservadas en todas nuestras estructuras, por lo que podemos concluir que la única diferencia entre estas estructuras es el posicionamiento de las cadenas laterales. Es decir, estamos evaluando diferentes conformaciones de la secuencia de la proteína silvestre.

Nuestros resultados indican precisamente este efecto, hay ciertas mutaciones que caen fuera del sitio de pegado que tienen resultados muy altos. Esto implica que existen

varios confórmeros de la secuencia silvestre con resultados muy altos y muy bajos. Estos resultados sugieren que la limitación que tenemos al hacer esta metodología es en el posicionamiento correcto de las cadenas laterales.

Debido a la gran cantidad de herramientas diferentes que se usan en esta metodología, hay múltiples factores que están afectando nuestros resultados y que vale la pena discutir:

1.-Prediccion de estructura: Cada vez que se evalúa una mutación, se genera una predicción de la estructura de dicha mutación, la cual se usa en el acoplamiento molecular. Esta predicción está hecha con el programa Modeller [13]. Este programa genera un alineamiento entre la secuencia del templado (la proteína silvestre) y la proteína con mutaciones. De acuerdo a nuestro análisis de mutaciones sencillas, nuestras proteínas a modelar tienen un 99% de identidad, debido a esto, la estructura predicha tiene los carbonos alfa en posiciones muy similares a la de la estructura silvestre.

Si usamos como premisa que las estructuras de las variantes y la proteínas silvestres van a ser sumamente cercanas, una mejora que se podría hacer es, usar la estructura de la proteína silvestre, y solo modificar dentro de esta estructura removiendo el aminoácido natural e insertando el mutante. De esta forma nos ahorramos calcular el posicionamiento del resto de las cadenas laterales, que mantendrá la posición silvestre.

2.-Posicionamiento de cadenas laterales: Nuestros resultados demostraron que estructuras con exactamente la misma secuencia de aminoácidos en la interfaz del complejo proteína-ligando tienen sus cadenas laterales posicionadas en lugares diferentes y esto a su vez genera una gran variabilidad dentro de los resultados durante el acoplamiento molecular. Este es un problema difícil de resolver, pero una posible solución sería intentar mejorar los parámetros usados en el paso de minimización.

3.-Número de soluciones a evaluar: Al momento de hacer el acoplamiento molecular, sólo usamos una estructura por mutación, a pesar de que Modeller tenga múltiples predicciones. Otra posible mejora seria usar múltiples predicciones para el acoplamiento molecular.

4.- Funciones que se usan para calificar nuestros resultados: Durante este estudio se decidió usar acoplamiento molecular basado en complementariedad geométrica debido a su gran rapidez y a la naturaleza del problema. Complementariedad geométrica ha sido muy útil para predecir los sitios de interacción entre proteínas, pero tal vez no tenga el nivel de resolución para predecir interacciones más sutiles como pueden ser puentes de hidrogeno. Usar otra función para evaluar nuestras mutaciones nos daría mayor resolución y una gran mejora en nuestra forma de calificar dichas soluciones; sin embargo, también aumentaría nuestro tiempo de cómputo en una manera exponencial.

5.-Una mutación puede requerir otras mutaciones para ser efectiva. Al generar una mutación en una proteína, puede que esta este teniendo interacciones con residuos físicamente cercanos. Debido a esto, puede que se requieran mutaciones adicionales para que esta mutación tenga el efecto deseado(por ejemplo que haya problemas de espacio en el sitio de interacción, y la cadena lateral de la mutación este teniendo choques con las demás cadenas).Debido a este efecto, muchas predicciones computacionales han sido acompañadas de rondas de evolución dirigida [2] o por rondas de optimización de estructura con ROSETTA [4].

6.- El cambio de cadenas laterales al interactuar con el ligando. Nosotros predecimos la estructura de nuestra mutación en ausencia del ligando y usamos esta estructura para el acoplamiento molecular. Esto es un problema, ya que la posición de las cadenas laterales del sitio de pegado en ausencia del ligando probablemente son diferentes a la posición de las mismas en presencia del ligando, y éste es un efecto que no estamos tomando en cuenta en el presente estudio.

En resumen, consideramos que la metodología de diseño de proteína planteada en la presente tesis se va a ver ampliamente beneficiada en un futuro próximo a medida de que se adquiera un mayor poder de cómputo y se obtenga un avance en las herramientas

de modelaje estructural y de acoplamiento molecular. Sin embargo por el momento, esta metodología no  tiene la capacidad de poder evolucionar una proteína exitosamente.

*Diseño de proteínas a partir de bibliotecas de rotámeros.*

En el grupo de diseño computacional de proteínas en el Instituto Max Planck de Biología del Desarrollo en Tuebingen, Alemania, se ha tenido como objetivo el desarrollo de una metodología para el diseño de proteínas usando como templado inicial la estructura de una proteína acoplada a su ligando. El objetivo de esta metodología es cambiar la especificidad de una proteína de su ligando por un nuevo ligando de geometría similar. Esta metodología sigue los siguientes pasos:

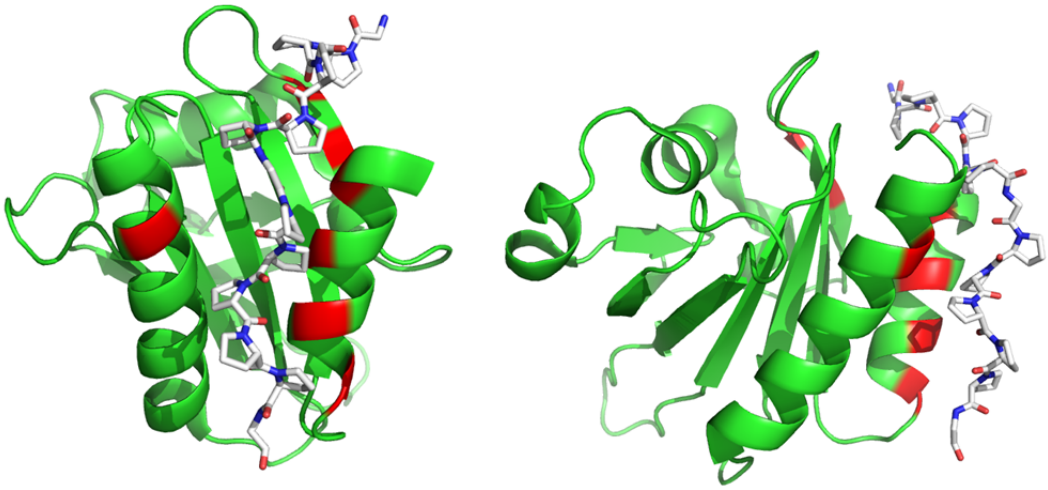


**Figura 16. Residuos de diseño.** En rojo se encuentran los residuos de "diseño", aquellos que van a ser mutados y que son los que tienen contacto directo con el ligando.

El primer paso es la selección de residuos de diseño(Figura 16); es decir los sitios que interactúan con el ligando natural y potencialmente con el nuevo ligando. Normalmente no se conoce *a priori* cuáles son estos residuos. Debido a lo anterior, normalmente estos residuos se escogen por distancia al ligando.

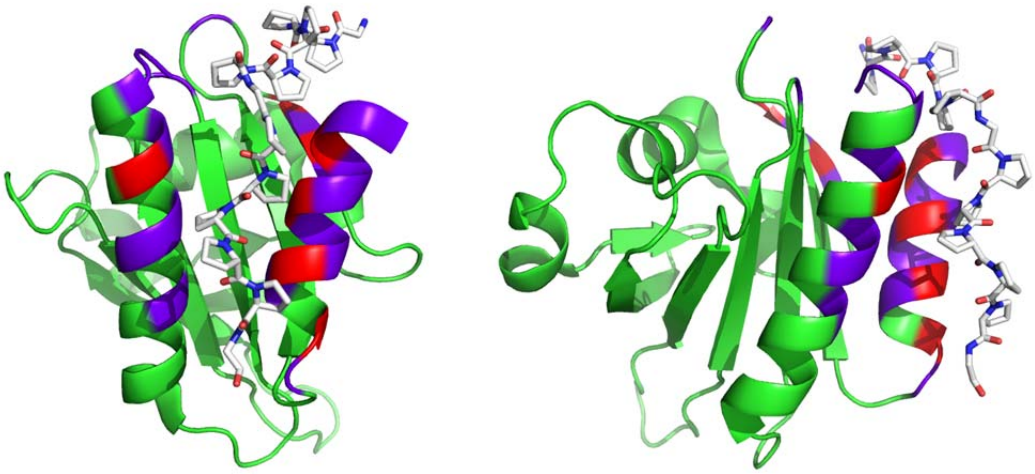


**Figura 17. Residuos flexibles.** En morado se encuentras los sitios flexibles, aquellos en los que únicamente se van a poder cambiar por rotámeros del mismo aminoácido. Esto es con el objetivo de que puedan acomodar las mutaciones en los residuos de diseño.

Una vez teniendo estos residuos, se escogen residuos "flexibles", estos son residuos cercanos a las posiciones de diseño. Se les llama "flexibles"(Figura 17) porque este algoritmo permitirá que sus cadenas laterales cambien y de esta forma, pueden acomodar las cadenas laterales de los residuos que sufrirán cambios. Normalmente estos residuos son aquellos que se encuentran cercanos a los residuos de diseño.

El segundo paso consiste en superponer el nuevo ligando a la posición del ligando natural en la estructura que ya se conoce, y usando estas coordenadas se generan múltiples confórmeros variando entre cada uno de ellos los 3 ejes (x, y, z) y las 3 posibles rotaciones (una por cada plano) de la posición del nuevo ligando. La Figura 18 es una representación virtual de los confórmeros del dominio rico en prolinas de VASP generado con esta permutación de rotaciones y cambios de coordenadas. Los confórmeros que choquen con el esqueleto de la proteína son descartados.

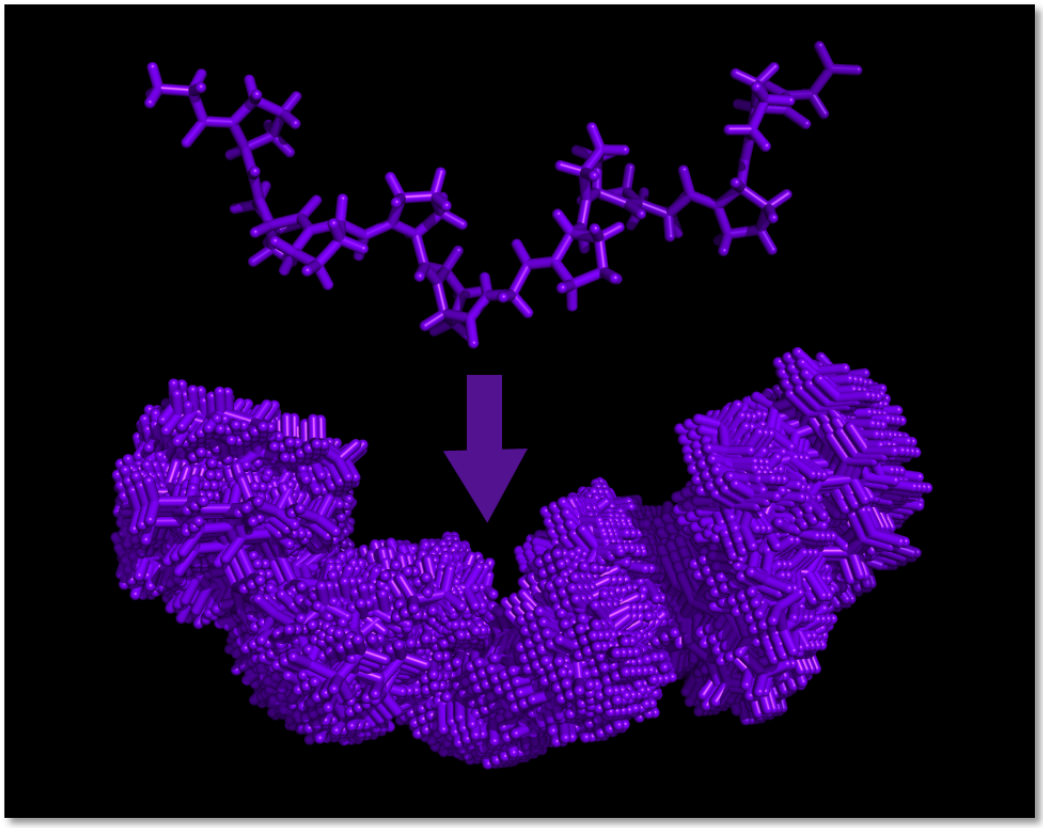


**Figura 18.Generacion de confórmeros a partir de la estructura de cristal del ligando.** Confórmeros generados haciendo rotaciones en todos los ejes (X, Y, Z) y translación en todos los ejes (X, Y, Z).

Habiendo escogido los residuos de diseño y los residuos flexibles, mediante el uso de una biblioteca de rotámeros obtenida del PDB, se calculan todas las siguientes energías. Existen diferentes funciones de energía que se pueden utilizar:

- La energía entre los átomos fijos y todos los confórmeros.
- La energía entre cada residuo y cada confórmero (las combinaciones que resulten en choques son descartadas)
- La energía entre cada rotámero y el esqueleto de la proteína.
- La energía entre cada par de rotámeros.

El programa busca la combinación de rotámeros que minimicen la energía total de la proteína, ésta incluye energía de empaquetado, y energía de estabilización con el

ligando. Para lograr esta tarea convierte todos los datos en un grafo dirigido, cada punto representa un confórmero, la conexión entre puntos tiene una distancia proporcional a la función de energía entre ese rotámero y el resto de la proteína y ligando. Lo que se busca es encontrar el camino más corto en este grafo; como cada punto representa un rotámero, y la distancia entre los rotámeros representa una función de energía, su busca un camino que pase para el menos un rotámero en cada aminoácido(para todo aminoácido debe de existir un rotámero) en que tenga la menor suma de energía, en este particular caso que tenga el camino mas corto. Este es un problema de optimización matemática que puede ser resuelto por optimización lineal, mejor conocida como programación lineal.

A pesar de que esta metodología está diseñada para trabajar con ligandos pequeños, debido a la rigidez de PLP, teóricamente podría usarse para optimizar profilina. Para explorar su potencial uso en este problema, se intentó mejorar la afinidad de profilina tipo II de ratón, con respecto al dominio rico en prolina VASP. Se ha hecho la caracterización de sitios funcionalmente importantes para el pegado a este dominio en esta variante de profilina. Como primer paso se escogieron como sitios de diseño estos sitios funcionalmente importantes.

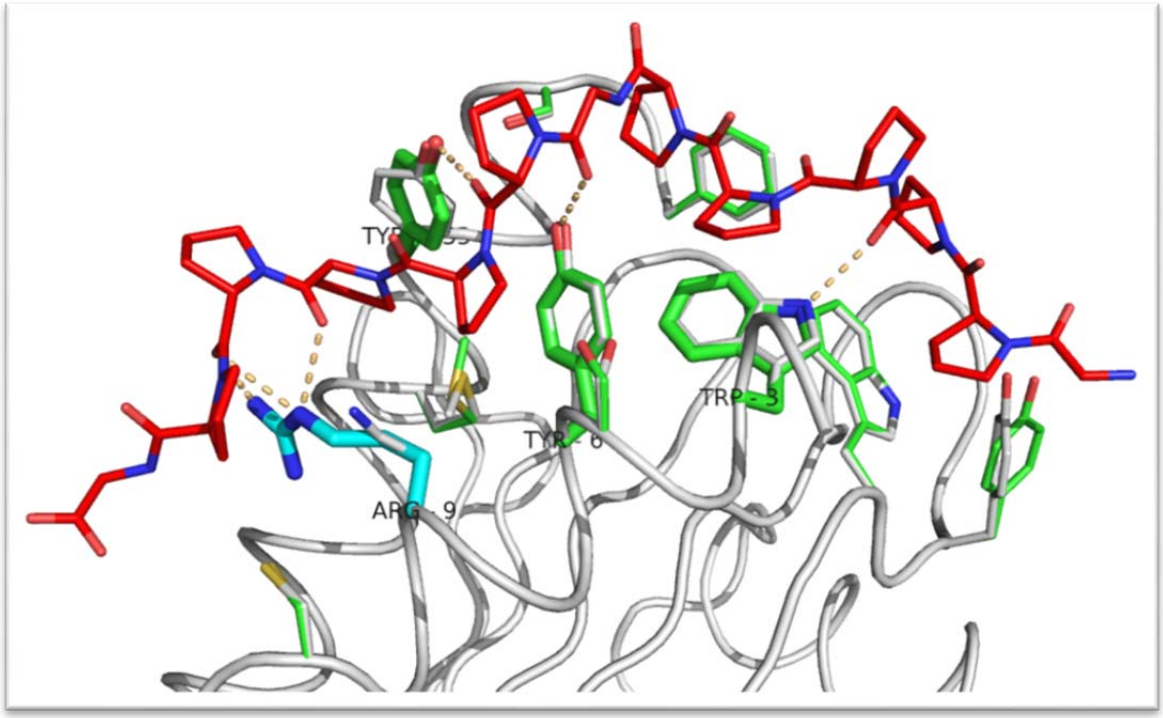


**Figura 19**. **Variante predicha de profilina** Una de las soluciones obtenidas mediante el diseño de proteínas usando bibliotecas de rotámeros.

Como podemos ver en los resultados, los residuos que fueron seleccionados como solución, son los residuos de la proteína silvestre a excepción de la Asp-9 por una Arg-9. También hay que notar que los rotámeros que fueron escogidos son muy parecidos a los rotámeros naturales.

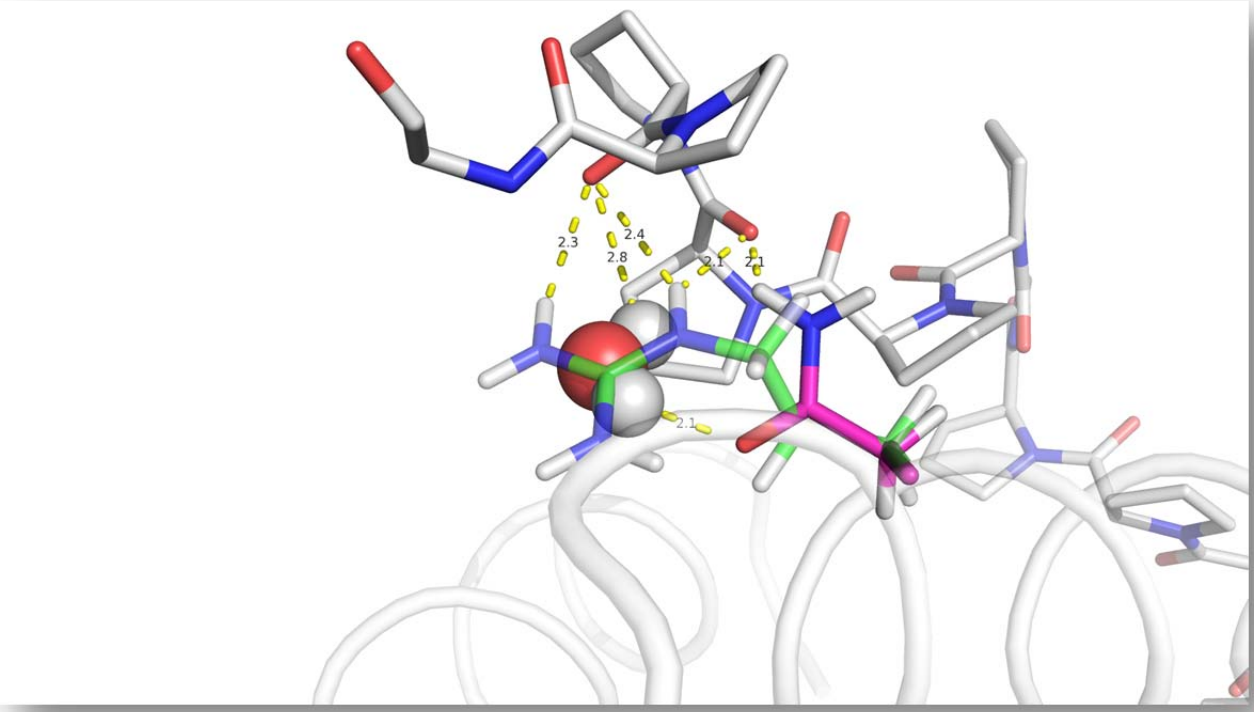



**Figura 20. Prediccion de interacciones electroestaticas entre cadenas**. Las marcas amarillas representan los puentes de hidrogenos predichos. En rossa se encuentra el residuo silvestre (ASP-9). En verde se encuentra el residuio de la mutacion (Arg)

Al observar más en detalle podemos ver que la estructura original tiene una Asparagina en el residuo 9 que está generando 2 puentes de hidrógeno; uno con PLP, y otro con una molécula de agua que a su vez está generando un segundo puente de hidrógeno con PLP. La mutación predicha por esta metodologia es una Arginina que está desplazando una molécula de agua y generando 3 puentes de hidrógeno directamente con PLP.

Una de las limitantes de esta metodología es el hecho de que no se puede escoger muchas posiciones de diseño, ya que el número de combinaciones posibles aumenta exponencialmente, por lo que se pueden probar una cantidad reducida de sitios. El segundo problema que tiene, es que una vez que se fijan las posiciones de los carbonos alfa estas no cambian en ninguna de las variantes. Esto es un problema para muchas proteinas ya que para el correcto pegado del ligando con la proteína la posición de los carbones alfa de la proteína sufren desplazamientos, y con esta metodología dichas

interacciones no podría ser modeladas. Hay otros programas que permiten el movimiento de los carbones alfa, pero aumenta exponencialmente el poder de computo necesario para correr la soluciones, entre los mas usados Rosetta.

A pesar de estas limitaciones, podemos ver que esta última metodología posicionó las cadenas laterales de sitios que han reportado ser esenciales para el pegado de PLP, y una mutación (R9N) que potencialmente podria resultar en el aumento de afinidad.

## *Bibliografía*


1. Marcotte, E. M. .: A combined algorithm for genome-wide prediction of protein function. Nature, 402, 83-86 (1999)

2. Siegel, J.: Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science(329), 309-313 (2010)

3. Mayo, S.: Achieving Stability and Conformational Specificity in Designed Proteins via Binary Patterning. J. Mol. Biol(305), 619-631 (2001)

4. Claren, J. .: Establishing wild-type levels of catalytic activity on natural and artificial (βα)8-barrel protein scaffolds. Proc Natl Acad Sci(106), 3704-9. (2009)

5. Valencia, A.: Critical Assessment of Techniques for Protein Structure Prediction. Curr. Op. Struct. Biol.(12), 368-373 (2002)

6. Wodak, S. J.: Computer analysis of protein-protein interaction. J. Mol. Biol.(124), 323-342 (1978)

7. Katchalski-Katzir E, S.: Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA(89), 2195–2199 (1992)

8. Kaufmann KW, L.: Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry(49), 2987-98 (2010)

9. T., H.: Probabilistic models and machine learning in structural bioinformatics. Stat Methods Med Res.(18), 505-26 (2009)

10. Pokala N, H.: Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility & specificity. J Mol Biol.(347), 203-227 (2005)

11. G. Guillen, L.: Biochemical Characterization of Profilin from Seeds of Phaseolus vulgaris. L. Plant Cell Physiol. (2001)

12. Thorn KS, C.: The crystal structure of a major allergen from plants. Structure 5, 19-32 (1997)

13. N. Eswar, M.: Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, 5.6.1-5.6.30 (2006)

14. Eric Bonabeau, M.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press,

New York (1999)

15. Mode, C.: Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science. InTech (2011)

16. Forrest, S.: Genetic algorithms: principles of natural selection applied to computation. Science 261, 872-878 (1993)

17. Rosaura Aparicio-Fabre, G.: Profilin tyrosine phosphorylation in poly-L-proline-binding. The Plant Journal(47), 491-500 (2006)

18. A. Fiser, R. K.: Modeling of loops in protein structures. Protein Science 9, 1753-1773 (2000)

19. Akella, P.: http://www.ecs.umass.edu/ece/labs/vlsicad/.