UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES
ACATLÁN

"PREDICCIÓN DE INTERFACES DE PROTEÍNAS MEDIANTE ANÁLISIS DE REDES "

T E S I S

QUE PARA OBTENER EL TÍTULO DE

Licenciado en Matemáticas Aplicadas y Computación

PRESENTA

Ricardo Corral Corral

Asesor: Dr. Gabriel del Rio Guerra

septiembre 2012

A mi mamá, por su confianza plena y apoyo incondicional.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por proporcionar fondos bajo el proyecto 82308-Q.

Al Dr. Gabriel del Rio Guerra por la guía en la realización de este trabajo, así como por proporcionarme espacio y recursos computacionales en las instalaciones del Instituto de Fisiología Celular, UNAM.

A la Dra. María Teresa Lara Ortiz por su asistencia en la gestión de trámites.

Índice general

"La única forma de conciliar el teorema de incompletitud de Gödel con la teoría cuántica sería admitir que toda teoría es completa y consistente al mismo tiempo, hasta que sea observada y se le obligue a ser o lo uno o lo otro."
— Ricardo Corral Corral.

Introducción

La predicción de sitios de unión en proteínas constituye un problema abierto en el área de la biología y su relevancia incluye aspectos básicos como aplicados. En particular, aún cuando hoy en día se conocen miles de sitios de unión en las proteínas, se desconocen las reglas que los definen. Poder identificar estos sitios aceleraría el desarrollo de nuevos fármacos dirigidos a inhibir estos sitios de unión y permitiría describir la función biológica de las proteínas de interes. Aunque existen diferentes metodologías computacionales que buscan predecir estos sitios de unión en proteínas, las más exitosas a la fecha son basadas en comparaciones, por lo que su contribución al entendimiento de las reglas que definen a los sitios de unión es muy limitada. En este trabajo se explora un enfoque *ab initio* que busca descifrar estas reglas basado en el análisis de las redes de interacción de residuos de las proteínas.

Antecedentes

La función de las proteínas está determinada por las interacciones que tienen con otras moléculas. Dada la relativamente limitada cantidad de información experimental de proteínas en complejo con sus interactores en las bases de datos actuales, la predicción de residuos de interfaz se ha convertido en una área activa de investigación en el campo de la bioinformática, como una manera de evaluar nuestro enendimiento de la función de proteínas [24]. Éstas predicciones están basadas en las características observadas

de los residuos de interfaz, tales como: conservación de secuencia, propensidades de los residuos en interfaces, propensidad de estructura secundaria, accesibilidad al solvente y entropía conformacional de las cadenas laterales [24]. Adicionalmente, un mejor desempeño en estas predicciones es obtenido mediante el aprendizaje automático en algun conjunto de datos mediante enfoques numéricos o probabilísticos.

Alternativamente, estudios previos sobre la dihidrofolato reductasa[23] y la TATA binding protein [31] han observado que los residuos de interfaz son centrales en sus redes de interacción de residuos, especialmente cuando éstos son detectados en ensambles de conformaciones de proteínas. En estos estudios previos, los ensambles conformacionales son usados para representar el hecho de que la unión con un ligando es un proceso cinético y consecuentemente puede involucrar múltiples conformaciones de la proteína. De hecho, las proteínas tienden a presentar diferentes conformaciones cuando éstas se encuentran en forma libre o en complejos.

Aún no existe evidencia estadística de que estas dos características (centralidad en las redes de interacción de residuos y los ensambles de conformaciones de proteínas) puedan predecir residuos de interfaz.

Además, estas características no son consideradas por los métodos actuales enfocados a predecir los residuos de interfaz de proteínas.

Objetivo

Analizar la eficiencia en la predicción de residuos de interfaz de proteínas mediante medidas de centralidad en redes de contacto y análisis de modos normales de proteínas.

Capítulo I

Marco teórico

I.1. Proteínas

Las proteínas son las macromoléculas biológicas más abundantes en las células. Existen miles de tipos de proteínas diferentes en una sola célula; muy distintas cada una en tamaño, peso molecular, y forma, lo que les confiere una enorme diversidad de funciones biológicas, y son la expresión final de toda la información genética.

La estructura de miles de proteínas diferentes, en organismos simples como bacterias o en formas de vida más complejas depende unicamente de 20 unidades simples llamadas aminoácidos. Estos aminoácidos se unen fuertemente (mediante enlaces químicos covalentes) en un orden particular.

Debido a que cada uno de estos aminoácidos tiene una cadena lateral que le confiere propiedades químicas características, la union de solo estos 20 aminoácidos en diferentes combinaciones puede dar origen a una gran variedad de proteínas como hormonas, enzimas, transportadores, anticuerpos, fibras musculares y muchas otras con actividades biológicas diferentes (Fig. I.1).

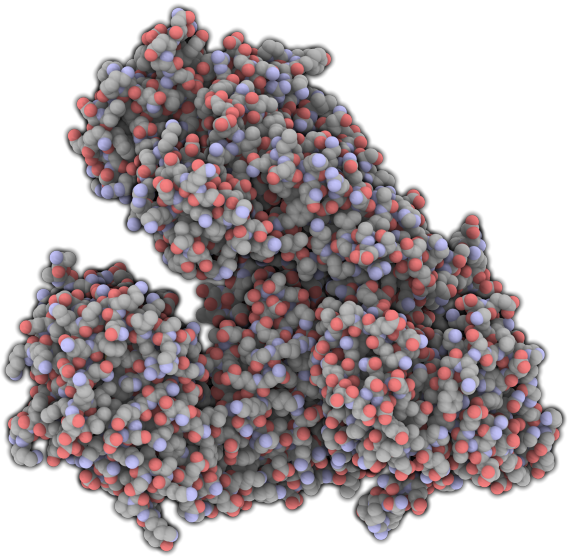

Figura I.1: Visualización generada por computadora del factor letal de la toxina Antrax mediante la aplicación QuteMol[37]. Las esferas representan una aproximación al espacio que ocupa cada átomo de carbono, hidrógeno y oxígeno

I.1.1. Aminoácidos

Una proteína es una serie de aminoácidos enlazados entre si mediante enlaces químicos fuertes; es decir, una proteína es un *polímero* de aminoácidos. Cuando un aminoácido es incorporado a una proteína se le denomina residuo de aminoácido, o simplemente residuo por la pérdida que sufre de elementos de agua al ser unido con otro. Cada residuo de aminoácido es unido con otro adyacente mediante un enlace químico covalente[10]. Es posible romper una proteína (hidrolizarla) por varios métodos y obtener de nuevo los aminoácidos independientes que la constituyen.
Estos aminoácidos comparten varias características estructurales. Todos es-

tan conformados por un grupo carboxilo y un grupo amino unidos a un mismo átomo de carbono denominado carbono alfa. Lo que los hace diferentes son sus cadenas laterales, o grupos R, los cuales varian en estructura, carga eléctrica y tamaño y tienen gran influencia en la solubilidad en agua. Las propiedades más generales de estos grupos se suelen clasificar principalmente en su polaridad, es decir, su tendencia a interactuar con el agua. Por ejemplo, aquellos no polares (hidrofóbicos), como los de la alanina, valina, leucina, e isoleucina tienden a aglutinarse en la proteína, dandole estabilidad a su estructura mediante interacciones hidrofóbicas.

Al enlace covalente que une un par de aminoácidos se le denomina enlace peptídico. A un par de aminoácidos unidos de esta manera se le llama dipéptido.

El enlace peptídico es formado mediante la eliminación de elementos de agua del grupo alfa carboxilo de un amino acido y el grupo alfa-amino de otro, como se ve en la figura I.2. Al polímero de varios aminoácidos se le denimina polipéptido.

Los términos polipéptido y proteína pueden ser usados indistintamente, pero generalmente se les nombra polipéptidos a moléculas de poco tamaño y proteína a las más grandes. Las proteínas pueden tener miles de residuos de aminoácidos.

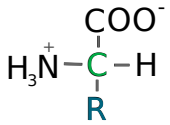

Figura I.2: Estructura general de un aminoácido. La cadena lateral R mostrada en azul, unida al carbón alfa, es distinta en cada aminoácido

I.1.2. Niveles de descripción estructural

La secuencia de aminoácidos codificada en el ADN determina a la proteína para la cual codifica y se le llama estructura primaria de la proteína.

Se postula[1] que la estructura tridimensional de la proteína depende directamente de su estructura primaria[9]. Esta configuración tridimensional otorga a la proteína sus características particulares. Se le llama estructura secundaria al arreglo tridimensional de la proteína en términos de conformaciones locales, tales conformaciones locales pueden adquirir un tipo particular de organización como hélices y hebras (Fig. I.3).
La interacción de estos elementos de estructura secundaria da origen a la estructura terciaria de la proteína (Fig. I.4).

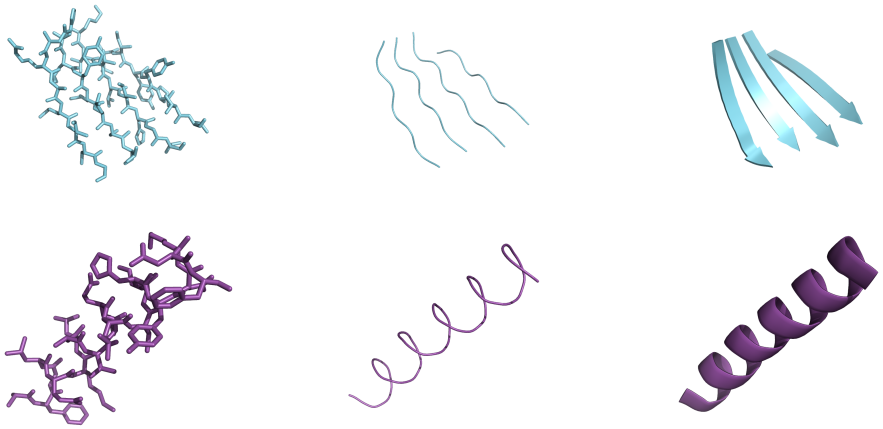

Figura I.3: Ejemplos de representaciones gráficas de elementos de estructura secundaria: hebras (arriba) y hélices (abajo), representadas de izquierda a derecha como enlaces atómicos, backbone y caricatura

[1]Este postulado es referido también como el dogma de Anfinsen o la hipótesis termodinámica del plegamiento.

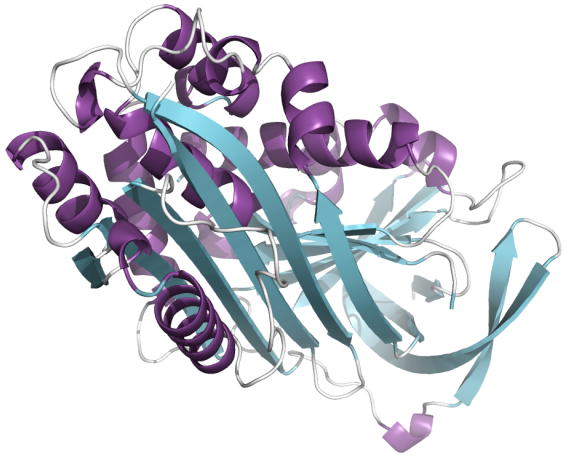

Figura I.4: Representación gráfica en caricatura de la estructura terciaria de una proteína completa con los elementos de estructura secundaria distinguibles.

I.2. Redes

El término red puede ser usado indistintamente por el de gráfica o grafo (del inglés *graph*) refiriendose a un objeto matemático construido a partir de elementos y relaciones entre ellos.
A continuación se daran algunas definiciones formales de estas gráficas y sus propiedades.
Sea V un conjunto finito, y denótese por

$$E(V) = \{\{u, v\} \mid u, v \in V, u \neq v\}$$

a los subconjuntos de un par de elementos distintos de V

Definición 1 *Un par $G = (V, E)$ con $E \subseteq E(V)$ es llamada gráfica en V. Los elementos en V son llamados vértices y los elementos en E aristas en G. El conjunto de vértices en una gráfica G se denota por V_G y su conjunto de aristas por E_G.*

A las gráficas también se les llama *gráficas simples* y a los vértices se les puede nombrar nodos.
Se escribirá la arista $\{u, v\}$ simplemente como uv . También por simplicidad se escribirá $v \in G$ y $e \in G$ en lugar de $v \in V_G$ y $e \in E_G$.
Dada una arista $e = uv$, se dice que e incide en u y en v.
Así, las aristas en una gráfica pueden describir cualquier tipo de relación entre los vértices. Se puede visualizar a los vértices como un conjunto de puntos y a las aristas como lineas que los unen (Fig. I.5).

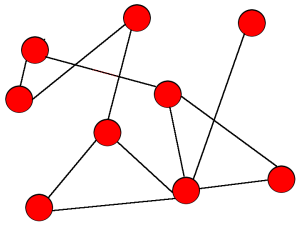

Figura I.5: Visualización de una gráfica G, con puntos rojos como vértices y lineas uniendo algunos pares de éstos como aristas.

En cada vértice pueden incidir un número variable de aristas. A éste número se le denomina grado.

Definición 2 *Sea $v \in G$ un vértice del grafo G, Los vértices adyacentes a v son el conjunto*

$$N_G(V) = \{u \in G \mid vu \in G\}$$

el grado de v es el número de vértices adyacentes

$$\delta_G(v) = \mid N_G(v) \mid$$

Un concepto fundamental en las gráficas es la conectividad, en su sentido más simple, cada arista uv representa la conectividad directa entre los vértices u y v. Este concepto se puede extender a la conectitud entre cualesquiera par de vértices en términos de alguna secuencia de aristas que unan vértices intermedios

Definición 3 *Sean $e_i = u_i u_{i+1} \in G$ aristas de G para $i \in [1, k]$. La secuencia $W = e_1 e_2 ... e_k$ es un camino de longitud k de u_1 a u_{k+1}*

Se escribirá simplemente $W : u_1 \xrightarrow{k} u_{k+1}$ para referirse a un camino de longitud k. Si existe un camino de u a v de alguna longitud, se escribirá $u \xrightarrow{*} v$.
Pueden existir varios caminos de un vértice a otro de distintas longitudes. Es conveniente distinguir aquel de longitud mínima.

Definición 4 *Si existe una ruta de u a v en G, sea*

$$d_G(u, v) = \text{mín}\{k \mid u \xrightarrow{k} v\}$$

la distancia de u a v.

Aquel camino $u \xrightarrow{k} v$ con $k = d_G(u, v)$ se le llamara la ruta más corta, o simplemente ruta de u a v.

Centralidad

Es común definir valores de centralidad para los vértices en términos de las rutas en las que pueda participar. Un valor de *centralidad* para un vértice describe su importancia relativa a los demás vértices en el grafo. Varias medidas de centralidad pueden capturar variaciones de la noción de importancia de un vértice.
Sea $\sigma_{st} = \sigma_{ts}$ el número de rutas de $s \in V_G$ a $t \in V_G$, donde $\sigma_{ss} = 1$ por convención, y sea $\sigma_{st}(v)$ el número de rutas de s a t en las que aparece algún vértice $v \in V_G$. Las siguientes son medidas de centralidad estándar:

- *closeness centrality*
$$C_C(v) = \frac{1}{\sum_{t \in V_G} d_G v, t} \tag{I.1}$$
- *graph centrality*
$$C_G(v) = \frac{1}{max_{t \in V_G} d_G v, t} \tag{I.2}$$

- *stress centrality*

$$C_S(v) = \sum_{s \neq v \neq t \in V_G} \sigma_{st}(v) \tag{I.3}$$

- *betweeness centrality*

$$C_C(v) = \sum_{s \neq v \neq t \in V_G} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{I.4}$$

Valores grandes de centralidad indican que un vértice puede alcanzar a otros en rutas relativamente cortas, o que el vértice se encuentra en una fracción considerable de las rutas que conectan a otros.

Subgráficas

Un subconjunto de vértices y aristas de una gráfica forman una *subgráfica*. Formalmente una gráfica H es una subgráfica de una gráfica G, denotada por $H \subseteq G$ si $V_H \subseteq V_G$ y $E_H \subseteq E_G$.
Una subgráfica $H \subseteq G$ es una subgráfica *inducida* si $E_H = E_G \cap E(V_H)$. Se dice que H es inducida por su conjunto V_H de vértices.
A cada subconjunto no vacío $A \subseteq V_G$ le corresponde una única subgráfica inducida

$$G[A] = (A, E_G \cap E(A)) \tag{I.5}$$

Clanes

Se dice que una gráfica es *completa* si existe una arista entre cualquier par de sus vértices. Una gráfica completa con n vértices se denota K_n.
Se le denomina *clan* de una gráfica G a una subgráfica $C \subseteq G$ completa. Si no existe un clan C' de G tal que $C \subseteq C'$, entonces C es un *clan maximal* [6, 41, 8].

Capítulo II

Fundamentos

Las actividades de las proteínas son mayormente dependientes de sus estructuras tridimensionales (geometría), de tal forma que cambios en sus conformaciones son frecuentemente ligados íntimamente a sus actividades[29, 17, 15]. Después de que las proteínas son primeramente sintetizadas por la maquinaria ribosomal, cambios conformacionales a gran escala tienen lugar para plegar la proteína en sus conformaciones nativas o funcionalmente activas. Seguido a esto, cambios conformacionales a menor escala son frecutemente necesarios por varias proteínas para llevar a cabo sus actividades. Las proteínas pueden cambiar entre diferentes conformaciones en respuesta a su ambiente para cumplir sus roles funcionales. En muchos casos, tales cambios conformacionales son el resultado de la unión de otras moléculas que regulan su actividad[11, 18].

Éstas características pueden ser modeladas como objetos y propiedades matemáticas que pueden someterse a cálculos computacionales.
El modelado de proteínas es un campo de rápido desarrollo que permite abordar varias preguntas biológicas; incluso observaciones obtenidas mediante modelado computacional han ayudado a entender aspectos de las proteínas y sus relaciones estructura-actividad[46]. Sin embargo esta relación estructura

actividad no resulta del todo clara aún y tampoco se cuenta con una forma de modelado del todo satisfactoria.

II.0.1. Red de contactos

La red de contactos se construye a partir de la geometría de una proteína. Ésta estructura tridimensional es descrita en archivos de texto plano en formato PDB[45], el cual contiene las coordenadas espaciales de los átomos que conforman cada residuo de aminoácido de la proteína. Ésta red de contactos toma como vértices a cada uno de estos átomos y define aristas sobre ellos dependiendo de interacciones entre los residuos (Fig. II.1). Para la construcción de la red de contactos en este trabajo, se toma en consideración exclusivamente la distancia euclidiana entre los átomos. Se define una arista entre cada par de residuos r_i y r_j, si cualquier átomo de r_i se encuentra a una distancia menor o igual a 5Å de cualquier átomo en r_j.

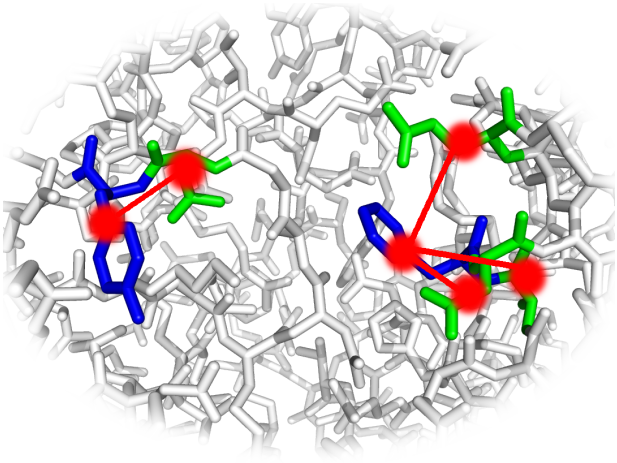

Figura II.1: Asignación de vértices y aristas a la red de contactos de una proteína. Considerando sólo los residuos de azul y verde, cada uno se toma como vértices en la red de contactos, ilustrados como puntos rojos. Las aristas, dibujadas como lineas rojas se establecen entre un par de residuos si éstos se encuentran cercanos en la estructura tridimensional de la proteína.

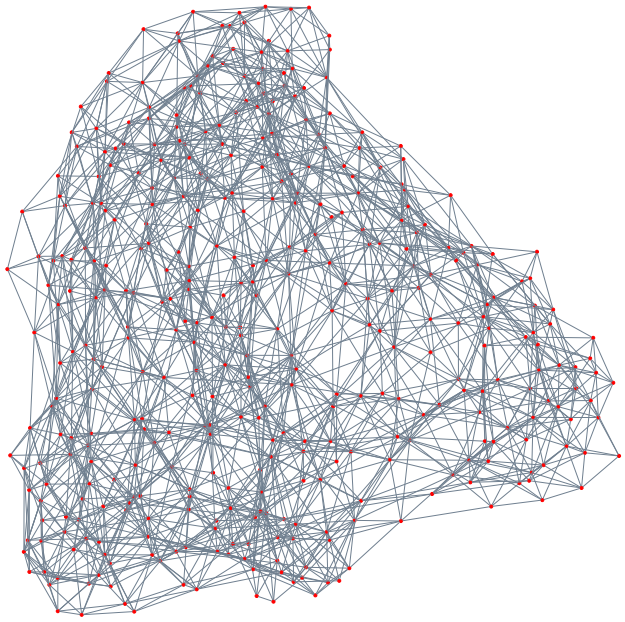

Figura II.2: Representación de la red de contactos de la proteína nauraminidasa N1 del virus de la influenza pandémica en 2009 H1N1. Cada uno de los 388 residuos es representado por un pequeño punto rojo. La red fue construida utilizando el algoritmo SFDP (Scalable Force Directed Placement)[22] para evitar la sobreposición de los vértices. La posición de los residuos en esta representación no corresponde con la localización en la estructura terciaria de la proteína.

Se denotará la red de contactos de una proteína P como $S_{RC}(P) = (V, E)$, con V vértices como residuos y E aristas.

II.0.2. Análisis de modos normales

El análisis de modos normales (NMA por sus siglas en inglés) es una herramienta utilizada para el estudio de movimientos colectivos en biomoléculas. Su uso ha llevado a la apreciación de la relación entre la estructura y comportamiento dinámico de las proteínas[32, 30, 3].

La aplicación más reciente del análisis de modos normales ha sido el predecir movimientos funcionales en proteínas y otras moléculas biológicas. Los movimientos funcionales son aquellos que se relacionan a su función y son causa o consecuencia de la unión con otras moléculas.
En estudios de NMA siempre se asume que los modos normales con mayor fluctuación (es decir, modos de frecuencia más baja) son aquellos funcionalmente relevantes[13].
NMA es un análisis armónico, es decir, es un análisis de moviemientos periódicos, oscilatorios y libres de fricción. Fundamentalmente utiliza los mismos campos de fuerza usados en simulaciones de dinámica molecular. En ese sentido es exacto (pues se asume que los campos de fuerza lo son), sin embargo, se sabe que la asunción de que la superficie de energía conformacional en un mínimo pueda ser aproximada por una parábola sobre un rango de fluctuaciones no es correcta a temperaturas fisiológicas[19, 34]. Existe bastante evidencia experimental y computacional de que la aproximación armónica falla para proteínas a temperaturas fisiológicas, donde lejos de efectuar un movimiento armónico en un único mínimo energético, el estado visita varios mínimos, atravesando barreras energéticas de distintas alturas. Asi que hay que estar consciente de esta asunción y sus limitaciones al efectuar NMA.
El NMA estándar requiere un conjunto de coordenadas, un campo de fuerzas que describa las interacciones entre los átomos constituyentes, y software que realize los cálculos requeridos.
La realización de un NMA en un espacio euclídeo requiere de tres pasos principales:

1. Minimización de la energía potencial conformacional como función de las coordenadas.

2. El cálculo de la matriz Hessiana, la cual es la matriz de segundas derivadas de la energía potencial con respecto a las coordenadas atómicas con pesos dependientes de su masa.

3. La diagonalización de la matriz Hessiana.

El último paso da como resultado valores y vectores propios (los "modos normales").
Cada uno de estos pasos puede ser demandante computacionalmente, sobre todo el primero y último. La minimización de energía requiere bastantes cálculos y por tanto utiliza tiempo de CPU y la diagonalización requiere tiempo en CPU y memoria por involucrar la diagonalización de una matriz de $3N \times 3N$, donde N es el número de átomos en la molécula.
Se le ha llamado NMA estándar a este análisis para distinguirlo del modelo de red elástica de NMA.

Modelo de red elástica

Debido a las dificultades computacionales del NMA estándar, actualmente el modelo de red elástica es muy popular. Este sigue siendo un NMA, pero el modelo de la proteína es simplificado drásticamente.
Como el nombre sugiere, los átomos son conectados por una red de aristas elásticas. Este método tiene principalmente un par de ventajas sobre el NMA estándar. La primera es que no hay necesidad de realizar la minimización de energía, ya que se considera que todas las conecciones elásticas se encuentran a su longitud de energía mínima. Y segundo, la tarea de diagonalización se reduce bastante comparada con el NMA estándar, pues se toma en cuenta el número de residuos en lugar de el número total de átomos si se consideran únicamente los átomos C^{α}.
A diferencia del NMA estándar, el modelo de red elástica requiere que se fijen un par de parámetros. Uno es la fuerza o constante de resorte denotada como γ y el otro es la distancia de corte R_c.
Se ha mostrado que hay bastante correspondencia entre el NMA estándar y el modelo de red elástica [40]. Dadas las asunciones drásticas que se hacen en el NMA estándar, la pequeña diferencia entre estos dos métodos es despreciable comparada con las diferencias entre el NMA estándar y la realidad.

Teoría de análisis de modos normales

La idea básica es que una función potencial V en un mínimo puede ser expandida en series de Taylor en términos de las coordenadas con pesos dependientes de la masa $q_i = \sqrt{m_i}\Delta x_i$, donde Δx_i es el desplazamiento de la coordenada i del mínimo energético y m_i es la masa del átomo correspondiente. Si la expansión se termina en el término cuadrático, entonces dado que el término lineal es cero en el mínimo energético , se tiene que:

$$V = \frac{1}{2}\sum_{i,j=1}^{3N} \left.\frac{\partial^2 V}{\partial q_i \partial q_j}\right|_0 q_i q_j \tag{II.1}$$

Así que la superfice de energía es aproximada por una parábola caracterizada por las segundas derivadas evaluadas en el mínimo energético.
Las segundas derivadas en (II.1) pueden ser escritas en una matriz F llamada comúnmente Hessiana. La determinación de sus valores y vectores propios (equivalente a la diagonalización) implica:

$$Fw_j = \omega_j^2 w_j \tag{II.2}$$

donde w_j es el j-ésimo vector propio y ω_j^2 su valor propio asociado. Cada vector propio especifica un modo normal mediante:

$$Q_j = \sum_{i=1}^{3N} w_{ij} q_i \tag{II.3}$$

La suma es sobre los elementos de w_j. Se puede mostrar que estos modos normales oscilan armónicamente e independiente de cada otro con la frecuencia angular ω_j^2:

$$Q_j = A_j cos(\omega_j t + \varepsilon_j) \tag{II.4}$$

Si un único modo normal j es activado, entonces:

$$\Delta x_{ij} = \frac{w_{ij}}{\sqrt{m_i}} A_j cos(\omega_j t + \varepsilon_j) \tag{II.5}$$

De manera que cada modo normal especifica un patrón de desplazamiento atómico.

Comúnmente se comparan los modos de frecuencia mínima con los modos funcionales, por ejemplo los derivados de un par de estructuras cristalográficas, en una unida a su ligando y en otra sin este.

No existe una diferencia esencial entre el NMA de red elástica y el NMA estándar además de el campo de fuerzas utilizado. En el caso del modelo de red elástica, el Hessiano puede ser derivado de la siguiente función potencial[40]:

$$V = \frac{\gamma}{2} \sum_{|r_{ij}^0|<R_c} \left(r_{ij} - r_{ij}^0\right)^2 \tag{II.6}$$

donde r_{ij} es la distancia entre los átomos i y j y r_{ij}^0 es la distancia entre los átomos en la estructura de referencia (la estructura cristalográfica por ejemplo). La constante de resorte γ es igual para todos los pares de átomos. La suma solo es efectuada sobre los átomos a una distancia de corte menor de R_c.

La red correspondiente a la función de energía de la ecuación (II.6) es algunas veces referida como el modelo de red anharmónica[7]. Una vez que la función de la ecuación (II.6) ha sido calculada se sigue el mismo procedimiento que en el NMA estándar. Se calcula el Hessiano y se determinan sus valores y vectores propios.

Mientras que en el NMA estándar se deben considerar todos los átomos, el modelo de red elástica puede utilizar solo un subconjunto de estos. Para una proteína por lo general son solo los C^α.

Comparado con el NMA estándar, esto puede resultar en un Hessiano de orden hasta diez veces menor, produciendo un gran ahorro computacional en el cálculo de los valores y vectores propios.

Capítulo III

Predicción de interfaz

III.1. Interfaces y sitios de unión

Las interacciónes que tiene una proteína con cualquier otra molécula está mediada principalmente por una fracción de sus residuos de aminoácidos. Frecuentemente se hace uso de los términos *sitio de unión* e *interfaz* para referirse a ellos.

El sitio de unión (tambien llamado centro de unión), son aquellos residuos de una proteína indispensables para la unión (binding) con su ligando, y son los principales responsables de la afinidad a este. Es decir, el sitio de unión es el conjunto de residuos con efecto funcional en unión. Si bien los residuos que afectan el plegamiento o estabilidad de la proteína pueden afectar la unión al ligando, no son estos los que se consideran importantes para la unión.

Una interfaz es la fracción de una proteína que se encuentra cercana con su interactor. La definición de una interfaz para una proteína se establece comúnmente por algún criterio fijo de *distancia de corte*. Así pues, aquellos residuos que se encuentren a no más de una distancia de corte del ligando conforman una interfaz.

La elección de esta distancia de corte no surge de ninguna propiedad intrínseca de la interacción proteína-ligando, por lo que se han propuesto definiciones

alternativas de interfaz [4] basados en criterios geométricos más sofisticados. Todos estos, sin embargo, dependientes en la idea de *cercanía* entre la proteína e interactor en cuestión.
Actualmente es ampliamente aceptado que los residuos responsables de la unión son aquellos que entran en contacto directo con su ligando. Es decir, se asume que el sitio de unión y la interfaz en una proteína son equivalentes. El mismo término sitio de unión refleja esta idea.

III.2. Bases para la predicción de interfaz

Varias características de las proteínas han sido relacionadas con las propiedades de las redes de interacción de residuos derivados de estas. Por ejemplo, la medida de centralidad *betweeness* (I.4) puede ser utilizada para identificar residuos importantes que actúen como centros de nucleación en el plegamiento de proteínas [44], o aquellos involucrados en interacciones proteína-proteína [1]. De igual forma, la centralidad *closeness* ($I.1$) sugiere residuos importantes para la función de la proteína [38], [14]. Por otra parte, propiedades de las redes de interacción como la conservación de cúmulos de residuos altamente conectados han sido asociadas a la estabilidad de la proteína [20] [26].

`JAMMING` [33] es capaz de identificar residuos críticos para la función de proteínas mediante análisis de centralidad (I.4). Un residuo crítico es aquel que al ser mutado afecta drásticamente la función de la proteína, ya sea como consecuencia de un plegamiento inadecuado, estabilidad y/o habilidad de unión al ligando reducidas.
Se ha observado que aquellos residuos críticos para el plegamiento se encuentran principalmente en el interior (core) de la proteína. Por lo tanto, una estrategia para identificar aquellos residuos críticos para unión es filtrar aquellos residuos expuestos en la superficie calculando su área de accesibili-

dad al solvente [25], lo cual es un filtro común usado por varios metodos para la predicción de interfaces.

Una vez filtrados aquellos residuos críticos accesibles al solvente de una proteína P, se puede centrar el análisis en una red $G_{sup}(P)$ que solo los considere a estos (Algoritmo .3.2).
Además, una vez conocido el conjunto de residuos críticos V_{crit} de P arrojado por `JAMMING`, se puede inducir I.5 una *Subred de Nodos Críticos* $CNS_{sup}(P) \subseteq G_{sup}(P)$, donde el conjunto E_{CNS} de aristas consiste en toda $e \in E_{G_{sup}}$ tal que $e \in E(V_{crit})$, es decir:

$$CNS_{sup}(P) = G_{sup}(P)[V_{crit}] = (V_{crit}, E_{G_{sup}} \cap E(V_{crit}))$$

Por otra parte, las interfaces constan de varios residuos localizados para interactuar con el ligando. La identificación de aglomerados de residuos puede sugerir tambien su participación en alguna interfaz. Estos cúmulos de residuos corresponden con el concepto de clan (ver página 18) de su red de interacción derivada. En particular, se propone que los clanes en $CNS(P)$ forman interfaces en P.

Para tomar en cuenta la naturaleza dinámica de las proteínas, la predicción de V_{crit} se realiza sobre un ensamble de conformaciones de una proteína generadas con el modelo de red elástica (ver fundamentos, página 24). El análisis de centralidad se realiza sobre cada conformación, de manera que V_{crit} resulta de la unión de los vértices centrales. Para la descripción en pseudocódigo del Algoritmo .3.1.

Hipótesis

> Los componentes más fuertemente conectados de la subred de nodos críticos externa de una proteína definen sus interfaces.

Los elementos necesarios para poder evaluar esta hipótesis deben incluir:

- Datos experimentales y estructurales de proteínas en complejo.
- Medidas estadísticas que permitan evaluar las predicciones.
- Un criterio que permita definir que es un sitio de unión a nivel estructural que coincida con datos funcionales.

III.2.1. Parámetros para la evaluación de la predicción

Dado que no se cuenta con un criterio para determinar la cantidad fija de elementos en cada interfaz, en la evaluación de las predicciones se introdujo el parámetro de simplicidad m.
Para un número de clan ω de CNS, se consideran únicamente los κ-clanes maximales que satisfagan la relación $\kappa \geq (\omega - m)$.

La definición de interfaz se establece a partir de una distancia de corte c de la proteína P con su ligando L. Dada una estructura tridimensional de P y L en complejo, se considera un residuos r_i de P como perteneciente a la interfaz, si cualquier átomo pesado[1] de r_i se encuentra a no más de una distancia c de cualquier átomo de L. Por tanto, se realizaron varias evaluaciones con distintos valores comúnmente usados de c para definir la interfaz observada.

[1]Se consideran átomos pesados a aquellos con masa atómica mayor a uno, es decir, se omite el Hidrógeno y se consideran el Carbono, Oxígeno, Nitrógeno, etc.

III.3. Predicción de interfaz

Se realizó la predicción de interfaz para las estructuras de proteínas en complejo con su ligando (ver métodos, pag. 59) y se evaluó la eficacia de las predicciones mediante el Coeficiente de Correlación de Matthews (ver métodos, pág. 36) para distintos valores de parámetros de simplicidad m y distancia de corte c.
Aunque se observó una correlación positiva entre las predicciones y las interfaces observadas (ver resultados, pág. 42) habia varios residuos pertenecientes a la interfaz no predichos (falsos negativos) y residuos predichos que no pertenecian a la interfaz (falsos positivos). Reflejado en valores bajos de especificidad y sensibilidad con intervalos de confianza para la media de [0.733-0.779] y [0.309-0.358] respectivamente.

III.3.1. Fuentes de error en las predicciones

La existencia de falsos positivos podría ser explicada por el valor rígido de distancia de corte c para definir a la interfaz observada. Es decir, un residuo predicho como perteneciente a una interfaz que se encuentre a una distancia $c + \epsilon$ del ligando, para cualquier valor de ϵ será evaluado como una falsa predicción. Siendo que un residuo cercano a la interfaz podría tambien estar interactuando con el ligando, a diferencia de uno localizado al extremo contrario de la proteína, por tanto, estos dos casos deberían ser ponderados de manera distinta.
Para tomar en cuenta este aspecto, se utilizó la puntuación de Binding-site Distance Test [35] para evaluar las predicciones sobre el mismo conjunto de datos. Bajo esta puntuación se califican positivamente aquellos residuos cercanos a la interfaz que antes eran considerados falsos positivos (ver métodos, pág.37).

Otra fuente de error en las predicciones es el uso de un solo ligando, ya

que es poco usual que una sola proteína pueda participar en una única interacción molecular. Se evaluaron los mismos complejos proteína ligando y se consideraron como verdaderos positivos en las predicciones aquellos residuos pertenecientes a cualquier interfaz entre las cadenas de la proteína (Figura V.6) con una mejora promedio en MCC de 0.052.

III.4. Hot-spot

Interacciones proteína-proteína juegan un papel fundamental en los procesos biológicos y son de gran importancia para las células vivas. Aunque los principios que gobiernan este proceso no son del todo entendidos, se sabe que la energía de unión no se encuentra uniformemente distribuida entre los residuos de interfaz, teniendo una gran contribución solo de un subconjunto pequeño. Estos residuos son referidos como hot-spots de unión.
Reciente interés en interfaces proteína-proteína como blancos farmacológicos ha enfatizado la importancia de identificar hot-spots sistemáticamente. Usualmente esto es hecho mediante experimentos de mutagénesis sitio-dirigida de sitio como la técnica de *alanine scanning*. Estos experimentos intentan evaluar el impacto en términos de la energía libre de unión causada por mutaciones a alanina de residuos específicos. Esto de cualquier manera puede demandar un esfuerzo experimental significativo. En este escenario, hay un creciente interés en predicción de hot-spots computacionalmente más rápida y más barata que pueda dirigir los esfuerzos experimentales solo en aquellos residuos que tengan más oportunidad de ser hot-spots[21].

III.4.1. Distribución de hot-spots en la superficie de proteínas

Los resultados de varios experimentos de alanine scanning para la determinación de hot-spots se encuentran accesibles en la base de datos AseDB[39]. Se encontró que un hot-spot puede o no pertenecer a una interfaz (ver re-

sultados, página 53). Es decir, su importancia para el proceso de unión al ligando no depende solo de su interacción directa con éste.

III.5. Predicción de hot-spots

El método descrito en este trabajo pretende identificar residuos importantes para unión sin impacto en plegamiento o estabilidad. Estos residuos o hot-spots no necesariamente se encuentran en una interfaz, por lo tanto, los resultados de las predicciones que resultan falsos positivos para la predicción de interfaz podrían ser verdaderos positivos para la identificación de hot-spots. Para probar esta idea se evaluaron las predicciones sobre las proteínas reportadas en AseDB (ver página 52).

Capítulo IV

Metodología

IV.1. Análisis de sensibilidad y especificidad

La sensibilidad y la especificidad son dos probabilidades comúnmente usadas para evaluar la eficacia de clasificadores binarios. Dado un grupo de objetos pertenecientes a un par de conjuntos, o clases mutuamente excluyentes, se define a un clasificador binario como cualquier método que prediga la partenencia de cada objeto a estas clases. Dada una clase C y un objeto ϱ, se define la clase observada de ϱ como $C_{obs}(\varrho) = C$ si $\varrho \in C$. La clase predicha de ϱ por el clasificador binario se define como $C_{pred}(\varrho)$
Cada predicción del clasificador puede pertenecer a una de cuatro categorías (Figura IV.1):

TP **ó verdadero positivo** Cuando el objeto ha sido predicho correctamente como perteneciente a una clase. $C_{obs}(\varrho) = C_{pred}(\varrho)$

TN **ó verdadero negativo** Cuando el objeto se ha clasificado como no perteneciente a una clase correctamente. $C_{pred}(\varrho) \neq C$ y $C_{obs}(\varrho) \neq C$

FP **ó falso positivo** Cuando el objeto no pertenece a la clase predicha. $C_{pred}(\varrho) = C$ y $C_{obs}(\varrho) \neq C$

FN **ó falso negativo** Cuando el objeto no se predice como no perteneciente a su clase. $C_{pred}(\varrho) \neq C$ y $C_{obs}(\varrho) = C$

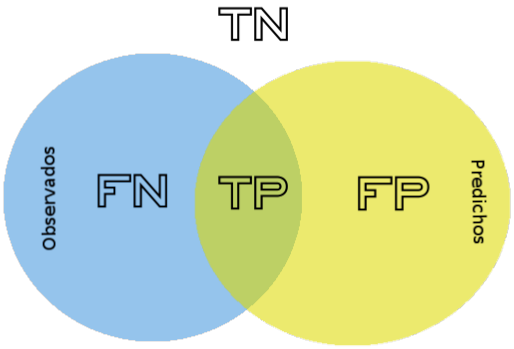

Figura IV.1: Diagrama de Venn de los conjuntos de predicciones y observaciones. Se ilustra la clasificación de cada región del diagrama como TP, FP, TN y FN.

La sensibilidad y especificidad estan dadas por las siguientes expresiones:

$$\textbf{Sensibilidad} = \frac{TP}{TP + FN}$$

$$\textbf{Especificidad} = \frac{TN}{TN + FP}$$

de manera que la sensibilidad es una medida de la habilidad del clasificador de identificar la pertenencia a una clase, y la especificidad de identificar la no pertenencia. Otra forma de tomar en cuenta los verdaderos y falsos positivos y negativos para evaluar la eficacia de un clasificador binario es el coeficiente de correlación de Matthews.

IV.2. Coeficiente de correlación de Matthews

El coeficiente de correlación de Matthews es un coeficiente derivado del coeficiente de correlación de Pearson para evaluar la calidad de clasificaciones binarias. Se calcula a partir de los valores de falsos y verdaderos positivos y negativos de la siguiente manera:

$$\mathbf{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad \text{(IV.1)}$$

Un valor del coeficiente igual a 1 representa una predicción perfecta, -1 una predicción inversa perfecta y 0 una predicción aleatoria. Su relación con los valores de sensibilidad y especificidad se puede observar en la figura IV.2 .

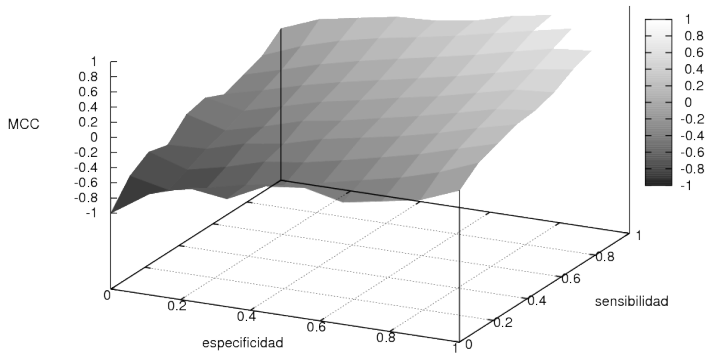

Figura IV.2: Interpolación de la superficie del coeficiente de correlación de Matthews como función de especificidad y sensibilidad para valores aleatorios de falsos y verdaderos positivos y negativos.

IV.3. Binding-site Distance Test

La puntuación del Binding-site Distance Test (BDT) es un método de evaluación de la precisión de predicciones de sitios de unión [35]. El Coeficiente de Correlación de Matthews (MCC) ha sido utilizado para evaluar la predicción de sitios de unión, tanto por desarrolladores de nuevos métodos, como por evaluadores en el experimento de predicción CASP (The Critical Assessment of protein Structure Prediction). Sin embargo, el MCC no toma en cuenta la posición tridimensional de los residuos predichos respecto a al interfaz observada. Además la puntuación obtenida medinte MCC depende mucho del criterio subjetivo (distancia de corte) con el cual se definen los residuos de unión observados. De manera que los residuos incorrectamente predichos obtienen exactamente la misma puntuación sin importar que tan cercanos se encuentran del sitio de unión *real*. El efecto en distintas predicciones en las puntuaciones MCC y BDT es ejemplificado en la figura IV.3 .
La puntuación BDT es calculada tomando en cuenta: La lista de los números de residuo que se predice que se unen al ligando, la lista de los números de residuos observados que se unen al ligando (interfaz), el archivo PDB de la estructura observada y una distancia de corte.

Se calcula la distancia euclidiana entre cada residuo del conjunto predicho y el observado y se transforma con la siguiente ecuación:

$$S_{ij} = \frac{1}{1 + \left(\frac{d_{ij}}{d_0}\right)^2}$$

Donde S_{ij} es la puntuación entre el residuo predicho i y el residuo observado j, d_{ij} es la distancia euclidiana entre los carbonos-α de los residuos i y j y d_0 es la distancia de corte.
La puntuación BDT final es simplemente la suma de las puntuaciones S_{ij} máximas de cada residuo predicho normalizada por el mayor valor del número

de residuos predichos (N_p) y el número de residuos observados (N_o):

$$BDT = \frac{\sum_{i=1}^{N_p} max(S_{ij})}{max(N_p, N_0)}$$

De esta manera, la puntuación BDT minimiza la penalización de residuos predichos ambiguos que podrian ser considerados como parte del sitio activo o en contacto con un ligando alternativo.

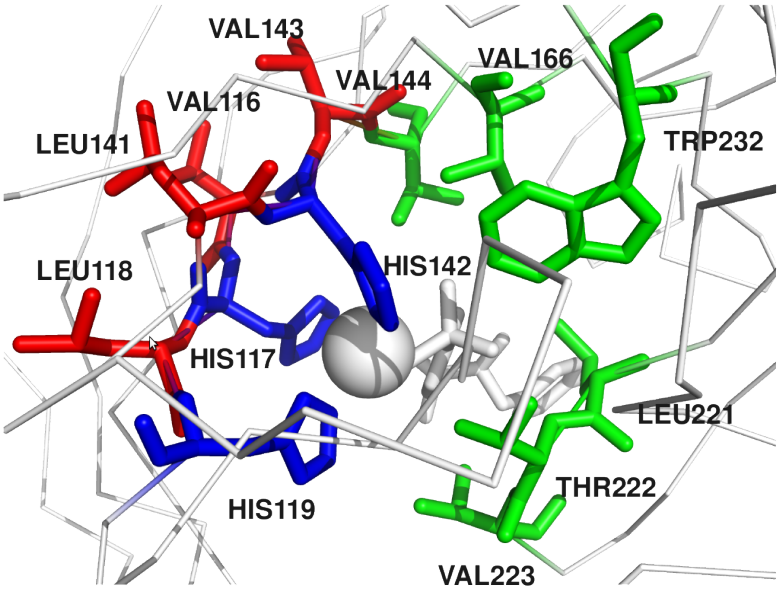

Figura IV.3: Diferencias entre BDT y MCC. En azul (residuos 117, 119, 142), la interfaz real, en rojo (residuos 116, 118, 141,143) y verde (residuos 114, 116, 221, 222, 223, 232) predicciones erroneas y neutras. Si la predicción neutra es tomada como TP, se tienen las puntuaciones MCC = 0.619 y BDT = 0.753, pero si es tomada como FP, se tienen MCC = -0.00001289 y BDT = 0.382. Si la predicción neutra no es tomada en cuenta en la evaluación se obtiene MCC = -0.0134 y BDT = 0.382.

Capítulo V

Resultados

V.1. Evaluación de predicción de interfaces

Se realizó la predicción de interfaces para las 6,412 proteínas en complejo con el ligando reportadas hasta marzo de 2010 en el Protein Data Bank. La eficacia de las predicciones con distintos valores de simplicidad y distancia de corte fué evaluada mediante el coeficiente de correlación de Matthews (Figuras V.1 V.2 V.3 V.4 V.6), sensibilidad y especificidad (Figura V.5) y la métrica del Binding-site Distance Test (Figura V.2).

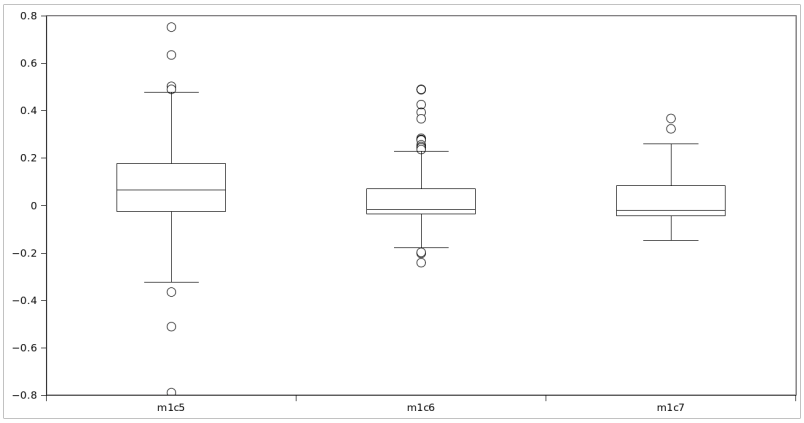

Figura V.1: Evaluación de predicciones de interfaz mediante el coeficiente de correlación de Matthews para el valor fijo de simplicidad $m = 1$ y distintas distancias de corte $c = 5, 6, 7$

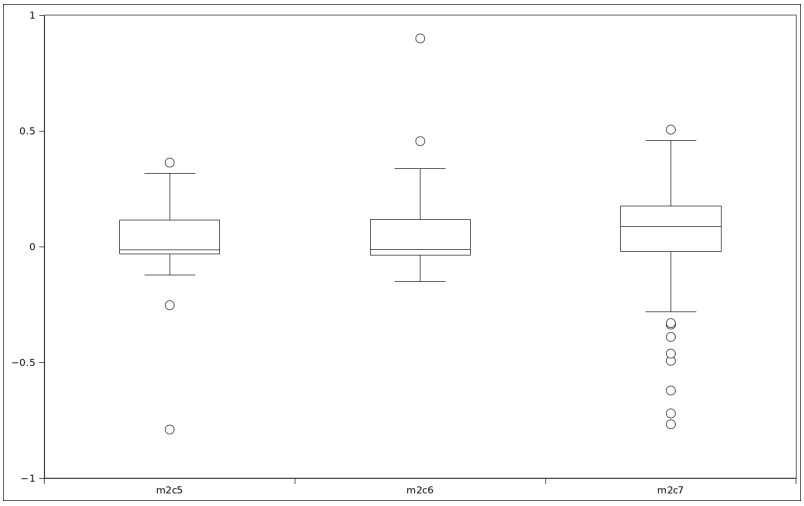

Figura V.2: Evaluación de predicciones de interfaz mediante el coeficiente de correlación de Matthews para el valor fijo de simplicidad $m = 2$ y distintas distancias de corte $c = 5, 6, 7$

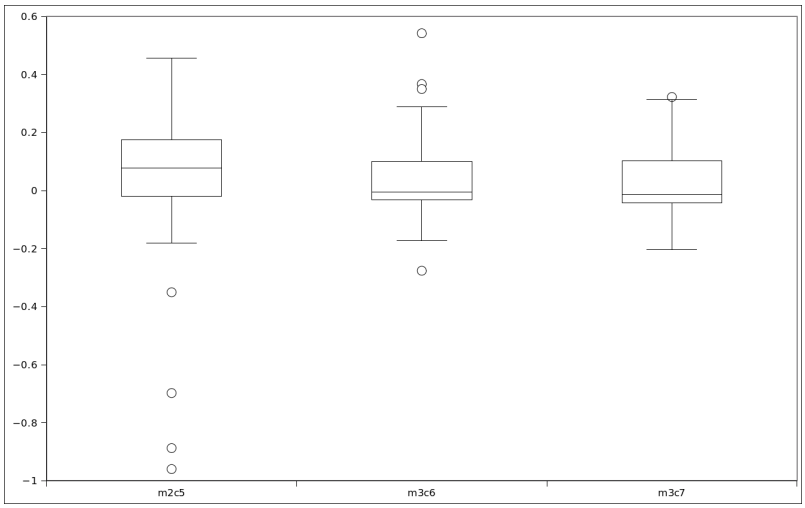

Figura V.3: Evaluación de predicciones de interfaz mediante el coeficiente de correlación de Matthews para el valor fijo de simplicidad $m = 3$ y distintas distancias de corte $c = 5, 6, 7$

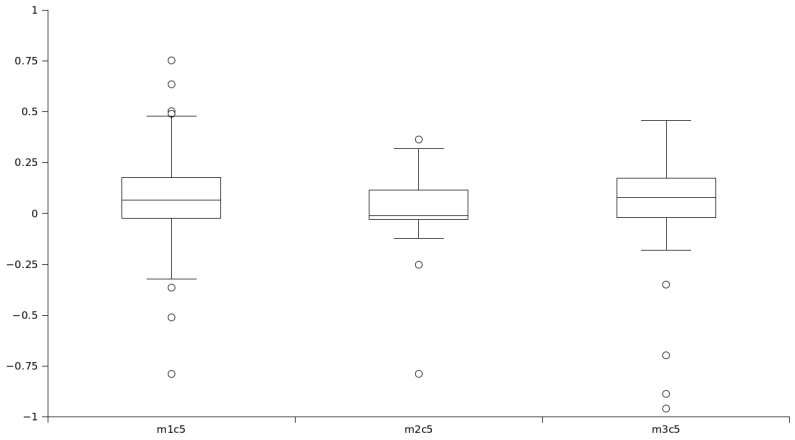

Figura V.4: Evaluación de predicciones de interfaz mediante el coeficiente de correlación de Matthews para distintos valores de simplicidad $m = 1, 2, 3$ y valor fijo de corte $c = 5$

Aunque el cambio en MCC para las predicciones con distintos valores de corte y simplicidad son muy similares, es más notorio el efecto en la sensibilidad y especificidad. En general las predicciones para interfaz son acompañadas por valores bajos de sensibilidad y altos de especificidad (Figura V.5).

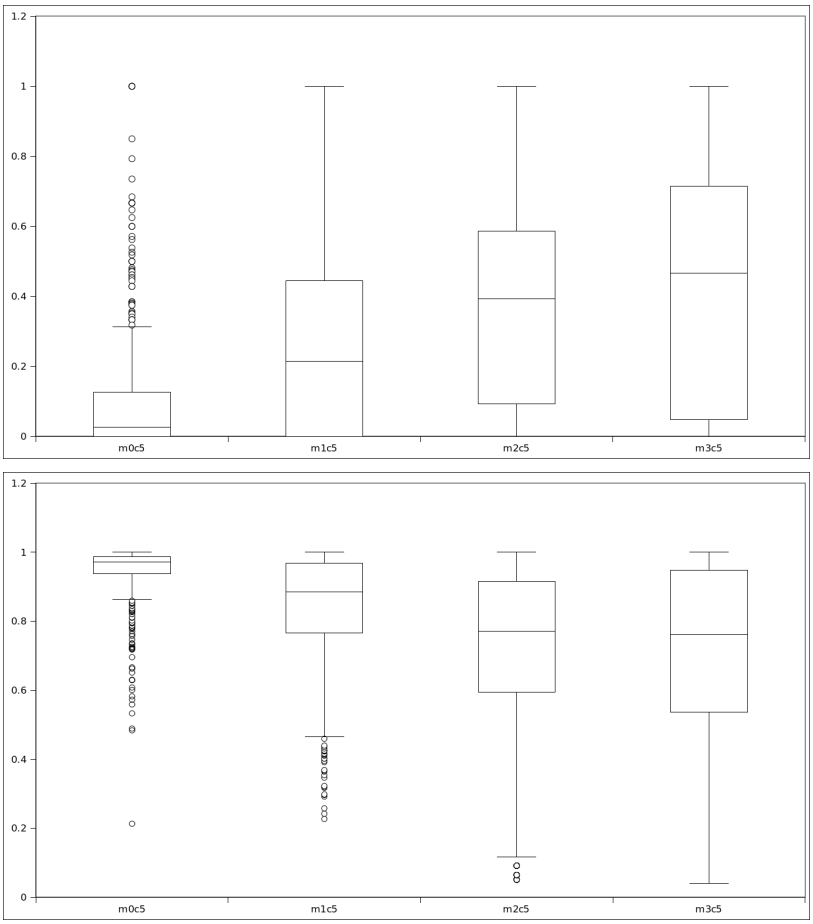

Figura V.5: Comparación de la sensibilidad (arriba) y especificidad (abajo) de las predicciones de interfaz para distintos valores de simplicidad $m = 0, 1, 2, 3$ y valor fijo de corte $c = 5$

La incorporación de residuos de interfaz entre las cadenas como parte de la interfaz observada resulta en una pequeña mejoría promedio en MCC de 0.052 (Figura V.6).

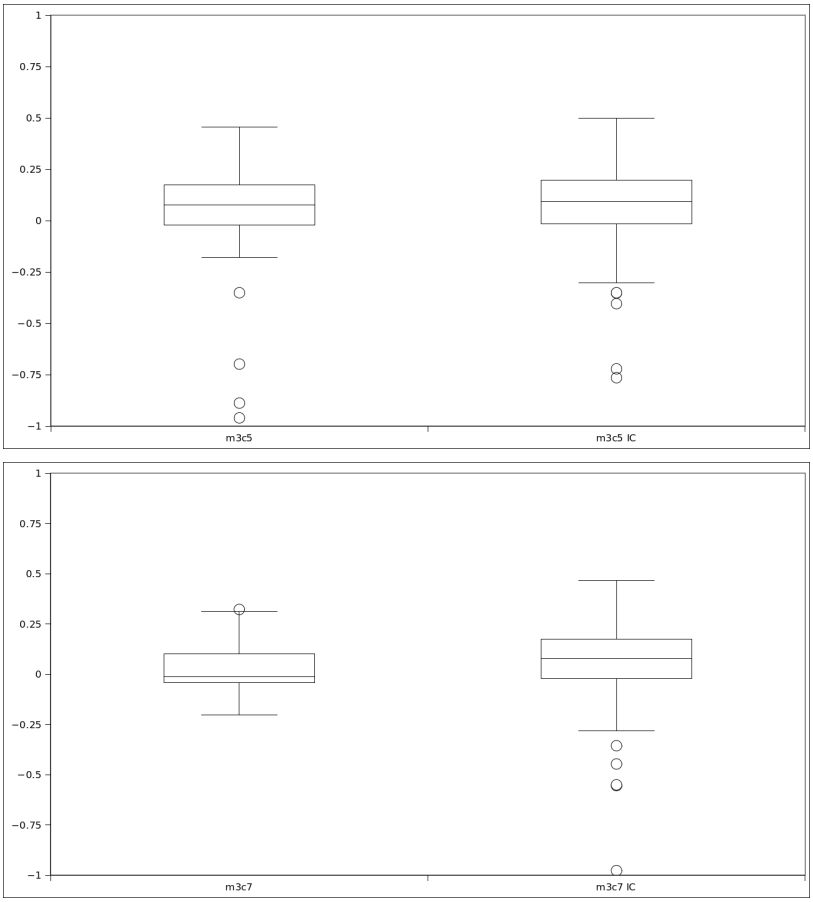

Figura V.6: Comparación del coeficiente de correlación de Matthews para las predicciones realizadas con y sin consideración de las interfaces entre cadenas de la misma proteína como interfaz observada a valores de corte de 5Å(arriba) y 7Å(abajo).

La evaluación de las predicciones de interfaz mediante BDT se muestran en la Figura V.7. La incorporación de interfaces entre cadenas a la evaluación de las predicciones resulta en una mejoría promedio en BDT de 0.22.

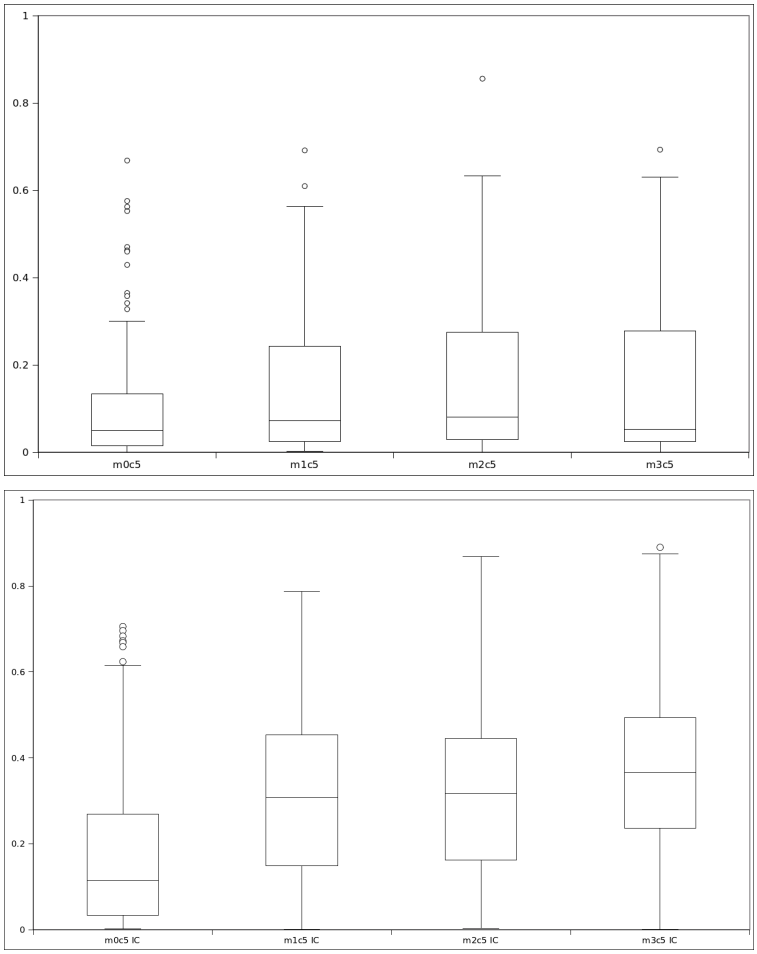

Figura V.7: Evaluación de predicciones de interfaz usando la puntuación de Binding-site Distance Test para distintos valores de simplicidad $m = 0, 1, 2, 3$ y valor fijo de corte $c = 5$. Tomando en cuenta sólo la interfaz con un ligando (arriba) y considerando interfaces entre cadenas (abajo).

V.2. Evaluación de predicción de hot-spots

Las predicciones de hot-spots se caracterizan por mayores valores en sensibilidad que especificidad (Figura V.8).

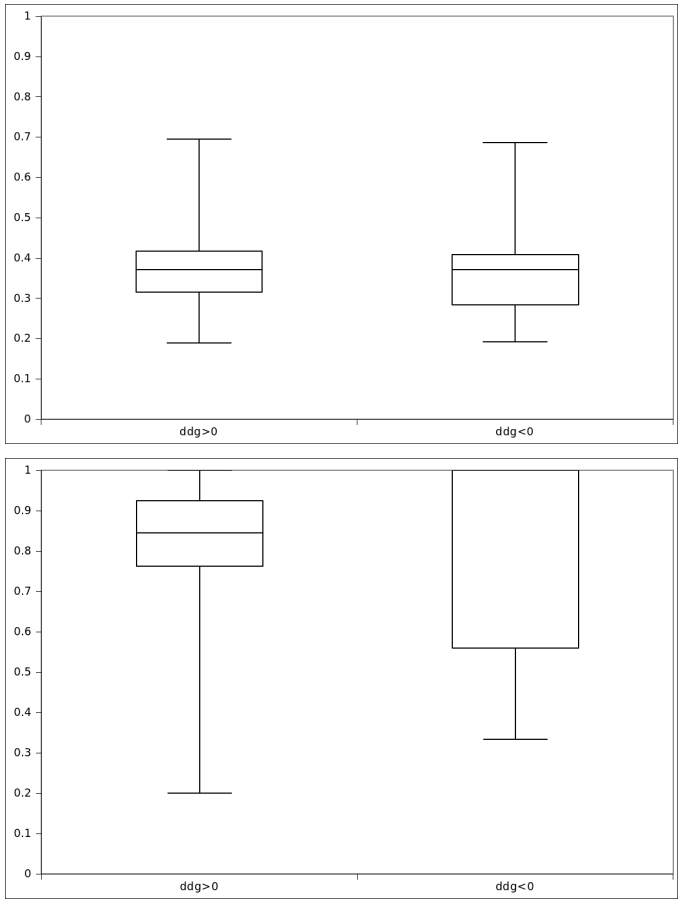

Figura V.8: Comparación de especificidad (arriba) y sensibilidad (abajo) de las predicciones de hot-spots con efecto negativo $\Delta\Delta G > 0$ y positivo $\Delta\Delta G < 0$ en unión.

Las estimaciones de los intervalos de confianza para la media de sensibilidad y especificidad se muestran en el cuadro V.1. Estos valores indican

una capacidad semejante de predecir reciduos cuya mutación tienen un efecto negativo en la unión al ligando ($\Delta\Delta G > 0$) o positivo ($\Delta\Delta G < 0$).
Puesto que no ha sido reportado algún predictor de hot-spots independiente de interfaz, estos valores no pueden ser directamente comparados, sin embargo, valores semejantes han sido reportados para la predicción de hot-spots dentro de interfaces basados en aprendizaje automático de parámetros estructurales y evolutivos con sensibilidad de 0.78 [21], y valores de accesibilidad al solvente y potenciales de contactos con una especificidad de 0.79 y sensibilidad de 0.59 [42], [43].

Los intervalos de confianza para la media de la evaluación BDT de predicción de interfaces y hot-spots se encuentran en el cuadro V.2.

	$\Delta\Delta G > 0$	$\Delta\Delta G < 0$
Sensibilidad	0.75-0.88	0.71-0.91
Especificidad	0.35-0.42	0.33-0.41

Tabla V.1: Intervalos de confianza para la media de sensibilidad y especificidad de las predicciones de hot-spots.

	BDT
Interfaz	0.37-0.38
Hot-spot	0.37-0.47

Tabla V.2: Intervalos de confianza para la media de BDT para predicciones de interfaz y hot-spot.

V.3. Distribución de residuos hot-spots dentro y fuera de interfaz.

Se muestran ejemplos (Figuras V.9 V.10 V.11 V.12) de la distribución de hot-spots fuera (en rojo) y dentro (azul) de interfaz de algunos de los complejos reportados en AseDB [39].

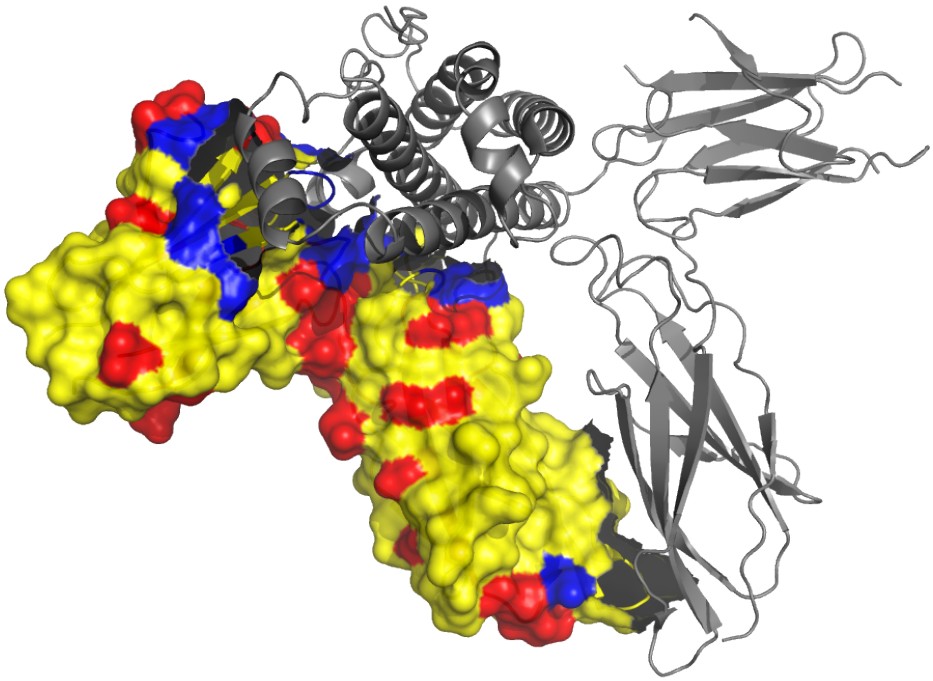

Figura V.9: 3HHR. Hormona de crecimiento humano y dominio extracelular de su receptor. De 61 residuos identificados como hot-spots [5] (Sensibilidad de predicción 0.82), solo 20 forman parte de una interfaz.

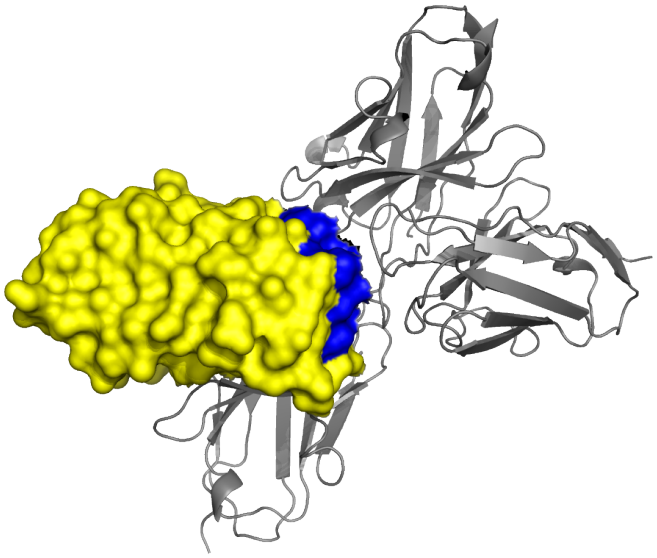

Figura V.10: 1DVF. Fv-Fv idiotope-anti-idiotope. Todos los hot-spots [16] forman parte de una interfaz.

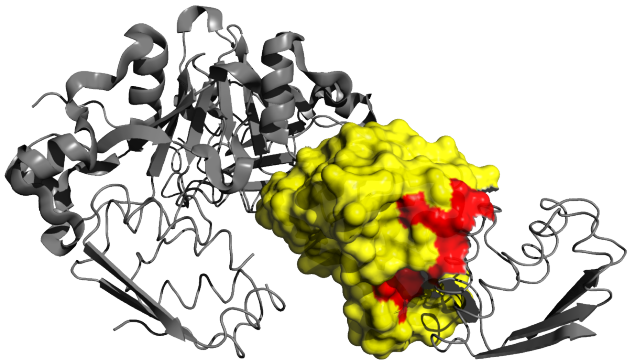

Figura V.11: 1BRS. Complejo Barnasa-Barstar. De todos los hot-spots [36] , ninguno forma parte de la interfaz si esta se define con la distancia de corte estándar de 5Ås.

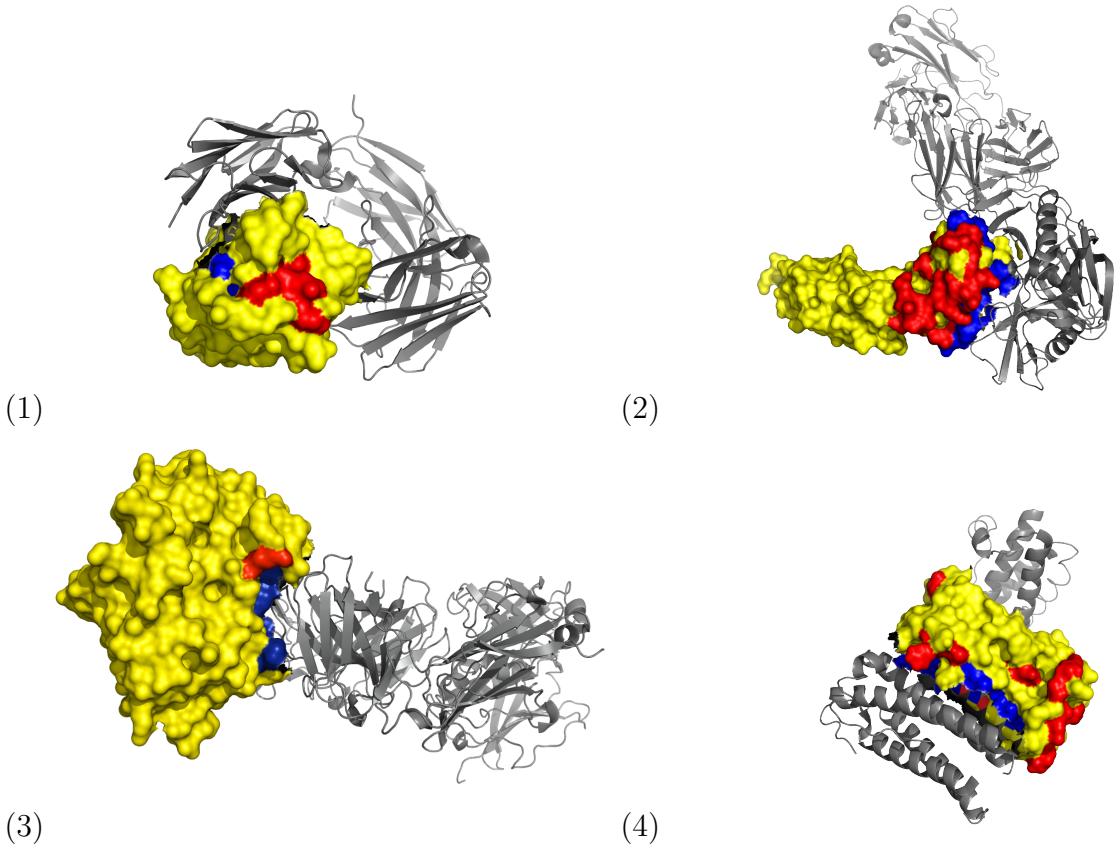

Figura V.12: Varios ejemplos de hot-spots e interfaces.
(1)1FVC: Fragmento IgG1-kappa del anticuerpo anti-p185HER2[27].
(2)1GC1: Receptor CD4 en complejo con la glicoproteína GP120 de la envoltura del virus HIV[2].
(3)1JCK: Enterotoxina estafilocócica C3, en complejo con un receptor de linfocitos T[28].
(4)1RHG: Factor estimulante de colonias de granulocitos[12].

Conclusiones

En resumen, los resultados obtenidos en este trabajo muestran que la geometría de proteínas (representadas mediante redes de interacción) puede ser efectivamente relacionada con los residuos en las interfaces de proteínas (es decir, su función) mediante la detección de residuos centrales en ensambles de conformaciones de proteínas derivadas mediante análisis de modos normales. De esta manera, la centralidad resulta ser una operación matemátca eficiente para relacionar la estructura con la función de proteínas. El método descrito es el primero en identificación de hot-spots dentro y fuera de interfaz y ofrece un desempeño semajante al de servidores web actuales dedicados a la predicción de hot-spots en interfaces de proteínas [21], [42], [43].

V.3.1. Discusión

El mismo análisis de redes descrito resulta ser capaz de identificar residuos fuera de interfaces pero importantes para la unión de la proteína con su ligando. En general, se desconocen los mecanismos por los cuales estos residuos fuera de la interfaz ejercen un efecto sobre la unión proteína-ligando. Por lo tanto, son necesarios más estudios sobre la relación de las propiedades de los modelos de proteínas basados en redes y los modelos bioquímicos de identificación e interacción molecular.

Materiales

.1. Sistema usado

Se utilizó un cluster CentOS 5.4 con 15 procesadores Opteron AMD Quad-Core 8380.
La mayoría de las aplicaciones fueron escritas en el lenguaje de programación C++. Para añadir la capacidad de cómputo paralelo con memoria compartida se usó OpenMP (http://openmp.org/). El código de las aplicaciones realizadas en C++ y en Fortran de los componentes de elNémo fueron compilados con `g++` y `gfortran` del GNU Compiler Collection (http://http://gcc.gnu.org/). La integración de las diferentes aplicaciones, asi como la implementación de algoritmos de cómputo no intensivo como la obtención de estadísticos para la evaluación de la eficacia de las predicciones fueron implementados en Python 2.7.1.

.2. Datos

Para la evaluación de la predicción de interfaces se utilizaron 6,412 estructuras de proteínas en complejo con el ligando reportadas hasta marzo de 2010 en el Protein Data Bank [45].
Los identificadores de las proteínas evaluadas reportadas en AseDB [39] son las siguientes: 5ebxA, 4fgfA, 3il8A, 3hhrC, 3hhrB, 3hhrA, 3hfmY, 3hfmH, 2ptcI, 2gf1A, 1vfbC, 1vfbB, 1vfbA, 1rhgC, 1rhgB, 1rhgA, 1rcbA, 1pneA,

1nt3A, 1jckD, 1jckC, 1jckB, 1jckA, 1gc1L, 1gc1H, 1gc1G, 1gc1C, 1fvcD, 1fvcC, 1fvcB, 1fvcA, 1dvfD, 1dvfC, 1dvfB, 1dvfA, 1danU, 1danT, 1danL, 1danH, 1cdcB, 1cdcA, 1cbwI, 1cbwD, 1bxiB, 1bxiA, 1brsF, 1brsE, 1brsD, 1brsC, 1brsB, 1brsA, 1ahwF, 1ahwD, 1ahwC, 1ahwA, 1a4yE, 1a4yD, 1a4yB, 1a4yA .

.3. Algoritmos

Calcula Interfaces .3.1

Entrada : Coordenadas atómicas de una proteína P, simplicidad m y número de confórmeros a generar z.
Salida : Conjunto de residuos de interfaz de P

Algorithm .3.1: CALCULA INTERFACES(P, m, z)

---

comment: Predice las interfaces de P

$MN \leftarrow$ MODOSNORMALES(P, z)
$interfaces \leftarrow \{\}$
$criticos \leftarrow \{\}$
for each $conformero \in MN$
  do $criticos \leftarrow criticos \cup$ JAMMING$(conformero)$
$G \leftarrow$ GSUP(P)
$CNS \leftarrow (V_G, \{(v_1, v_2) \in E_G : v_1, v_2 \in criticos\})$
$CM \leftarrow$ MAXCLIQUES(CNS)
$\omega \leftarrow max\{\|CM_i\|\}$
for each $clique \in CM$
  do $\begin{cases} \textbf{if } \|CM_i\| \geq (\omega - m) \\ \quad \textbf{then } interfaces \leftarrow interfaces \cup CM_i \end{cases}$
return $(interfaces)$

Algorithm .3.2: GSUP(P, d)

comment: Construye la red de contactos superficial de P

$V =\leftarrow \{\}$
$E =\leftarrow \{\}$
for each $residuo \in P$
  do if ASA$(residuo) \leq 6$
  then $V \leftarrow V \cup residuo$
for each $residuo_1 \in V$
  do { **for each** $residuo_2 \in V$
    do { **if** $residuo_1 \neq residuo_2$
      then for each $atomo_1 \in residuo_1$
        do { **for each** $atomo_2 \in residuo_2$
          do if $|atomo_1 - atomo_2| \leq d$
          then { $E \leftarrow E \cup \{residuo_1, residuo_2\}$
                    $continue$
return (V, E)

Nota : Éste pseudocódigo es sólo ilustrativo. Su complejidad es $\mathcal{O}\left(|V|^2\right)$. La organización de los puntos en estructuras de datos de particionamiento de espacio como los *árboles k-d* (Maneewongvatana y Mount, 1999) permiten la obtención de todos los puntos a no más de cierta distancia de otro en tiempo subcuadrático, permitiendo una complejidad final de $\mathcal{O}\left(|V| log^2\left(|V|\right)\right)$.

Bibliografía

[1] A . del Sol, H Fujihashi, and P OMeara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315, 2005.

[2] Ashkenazi A., Presta L. G., Marsters S. A., Camerato T. R., Rosenthal K. A., Fendly B. M., and Capon D. J. Mapping the cd4 binding site for human immunodeficiency virus by alanine-scanning mutagenesis. *Proc. Natl. Acad. Sci. USA*, 87:7150–7154, 1990.

[3] Brooks B. and Karplus M. Harmonic dynamics of proteins: Normal modes and fluc- tuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA*, 80:6571–6575, 1983.

[4] Y. E. Ban, H. Edelsbrunner, and J. Rudolph. Interface surfaces for protein-protein complexes. *J. Assoc. Comput. Mach.*, 53(3):361–378, 2006.

[5] S. H. Bass, M. G. Mulkerrin, and J. A. Wells. A systematic mutational analysis of hormone-binding determinants in the human growth hormone receptor. *Proc. Natl. Acad. Sci.*, 88:4498–4502, 1991.

[6] Bron C. and Kerbosch J. finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.

[7] Chennubhotla C., Rader A. J., Yang L. W., and Bahar I. Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. *Physical Biology*, 2:S173–S18, 2005.

[8] F. Cazals and C. Karande. A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, 407:564–568, 2008.

[9] Anfinsen CB. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[10] Thomas E. Creighton. *Proteins: Structures and Molecular Properties*. W. H. Freeman, second edition, 1992.

[11] B. L. de Groot, G. Vriend, and H. J. C. Berendsen. Conformational changes in the chaperonin groel: New insights into the allosteric mechanism. *J. Mol. Biol.*, 286:1241–1249, 1999.

[12] Reidhaar-Olson J. F., Souza-Hart J. A., and Selick H. E. Identification of residues critical to the activity of human granulocyte colony-stimulating factor. *Biochemistry*, 35:9034–9041, 1996.

[13] Tama F. and Sanejouand Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14:1–6, 2001.

[14] Amitai G. Network analysis of protein structures identifies functional residues. *J Mol Biol*, 344(4):1135–1146, December 2004.

[15] U. Gether, P. H. Andersen, and O. M. Larsson. Neurotransmitter transporters: molecular function of important drug targets. *Trends Pharmacol. Sci.*, 27:375–383, 2006.

[16] E. R. Goldman, W DallÁcqua, B. C. Braden, and R. A. Mariuzza. Analysis of binding interactions in an idiotope-antiidiotope protein-protein complex by double mutant cycles. *Biochemistry*, 36:49–56, 1997.

[17] L. Guan and H. R. Kaback. Lessons from lactose permease. *Annu. Rev. Biophys. Biomol. Struct.*, 35:67–91, 2006.

[18] K. Gunasekaran and R. Nussinov. How different are structurally flexible and rigid binding sites? sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J Mol Biol*, 365(1):257–273, 2007.

[19] Austin R. H., Beeson K. W., Eisenstein L., Frauenfelder H., and Gunsalus I. C. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14:5355–5373, 1975.

[20] J Heringa, P Argos, Egmond MR, and J de Vlieg. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng*, 8(1):21–30, 1995.

[21] Roberto Hiroshi Higa and Clésio Luis Tozzi. Prediction of binding hot spot residues by using structural and evolutionary parameters. *Genetics and Molecular Biology*, 32(3):626–633, 2009.

[22] Yifan Hu. Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, 2005.

[23] Z Hu, D Bowen, Southerland WM, A Del Sol, and Y Pan. Ligand binding and circular permutation modify residue interaction network in dhfr. *PLoS Comp Biol*, 2, 2007.

[24] Zhou H.X. and S Quin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23:2203–2209, 2007.

[25] Kabsch W and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[26] Brinda KV and S Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophys J*, 89(6):4159–4170, 2005.

[27] Cedergren L., Andersson R., Jansson B., Uhlen M., and Nilsson B. Mutational analysis of the interaction between staphylococcal protein a and human igg1. *Prot. Eng.*, 6:441–448, 1993.

[28] L Leder, Llera A., Lavoie P. M., Lebedeva M. I., Li H., Sekaly R. P., Bohach G. A., Gahr P. J., Schlievert P. M., Karjalainen, and Mariuzza R. A. A mutational analysis of the binding of staphylococcal enterotoxind b and c3 to the t cell receptor beta chain and major histocompatibility complex class ii. *J. Exp. Med.*, 187:823–833, 1998.

[29] Lemieux, Huang Y M. J., and D. N. Wang. The structural basis of substrate translocation by the escherichia coli glycerol-3-phosphate transporter: a member of the major facilitator superfamily. *Curr. Opin. Struct. Biol*, 14:405–412, 2004.

[30] Levitt M., Sander C., and Stern P. S. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.*, 10:181–199, 1983.

[31] Héctor Marlosti Montiel Molina, César Millán-Pacheco, Nina Pastor, and Gabriel del Rio. Computer-based screening of functional conformers of proteins. *PLoS Comput Biol*, 4(2), 2008.

[32] Go N., Noguti T., and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80:3696–3700, 1983.

[33] Michael P., Thibert Boris, Bredesen Dale E., and del Rio Gabriel Cusack. Efficient identification of critical residues based only on protein structure by network analysis. *PLoS ONE*, 2(5):e421, 2007.

[34] Elber R. and Karplus M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science*, 235:318–321, 1987.

[35] Daniel B. Roche, Stuart J. Tetchner, and Liam J. McGuffin. The binding site distance test score: A robust method for the assessment of predicted protein binding sites. *Bioinformatics*, 26(22):2920–2921, 2010.

[36] G. Schreiber and A. R Fersht. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.*, 248:478–486, 1995.

[37] Marco Tarini, Paolo Cignoni, and Claudio Montani. Ambient occlusion and edge cueing for enhancing real time molecular visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1237–1244, 2006.

[38] B Thibert, Bredesen Dale E., and Gabriel del Rio. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, 6(213), 2005.

[39] Kurt S. Thorn and Andrew A. Bogan. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3 2001):284–285, 2000.

[40] M.M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev.*, 77:1905–1908, 1996.

[41] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42, 2006.

[42] Nurcan Tuncbag, Attila Gursoy, and Ozlem Keskin. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520, 2009.

[43] Nurcan Tuncbag, Ozlem Keskin, and Attila Gursoy. Hotpoint: hot spot prediction server for protein interfaces. *Nucleic Acids Research*, 38:W402–W406, 2010.

[44] M Vendruscolo, Dokholyan NV, E Paci, and M Karplus. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65(6 Pt 1), 2002.

[45] J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[46] David S. Wishart. *Identifying Putative Drug Targets and Potential Drug Leads Starting Points for Virtual Screening and Docking*, volume 443, pages 333–351. Humana Press, 2008.