Universidad de Concepción

Facultad De Ingeniería.

Departamento de Ingeniería Informática
y Ciencias de la Computación

Algoritmo de ranking para complejos de proteínas basado en características fisicoquímicas

Autor:
Felipe E. Valenzuela Figueroa

Comisión:
Dr. Pedro Pinacho Davidson

Patrocinante:
Dra. Cecilia Hernández Rivas

Comisión:
Dr. Javier Vidal Valenzuela

Memoria de título presentada a la facultad de ingeniería para optar al título profesional de ingeniero civil informático.

16 de Diciembre, 2018

Resumen

Identificar y describir la interacción entre proteínas es un problema ampliamente estudiado. Si un grupo de proteínas trabaja en conjunto es llamado complejo de proteínas (PC), y regularmente está asociado a una función específica. Existen herramientas computacionales que predicen complejos de proteína desde muchos enfoques.

El resultado de estos predictores de complejos puede ser validado utilizando los Gold Standard Dataset (GSD), que son catálogos de referencia que incluyen información de proteínas experimentalmente aisladas, estudiadas y documentadas. De esta validación los resultados del predictor se clasifican como verdaderos positivos (VP) cuando están descritos en los GSD ó falsos positivos (FP) cuando no lo están.

Un resultado del predictor clasificado como FP puede interpretarse como un complejo no descubierto o con poca documentación experimental, por ese motivo, los FP son un grupo interesante de estudio. El problema, es muy costosa la validación experimental de un complejo, por eso evaluar experimentalmente los FP debe planificarse y no hacerse aleatoriamente.

Este trabajo propone un algoritmo de ranking para resultados FP basado en técnicas como el alineamiento estructural múltiple de proteínas (MPStrA) y el cálculo de índices fisicoquímicos de las proteínas. La función de orden del ranking busca dar el primer lugar de la lista al resultado FP que tenga más probabilidad de ser complejo, luego al segundo lugar más probabilidad que el tercero y así sucesivamente.

Los resultados del algoritmo fueron exitosos al identificar resultados FP que fueron documentados como complejos posterior al año de la referencia de proteínas usada.

Índice

Índice de figuras

Índice de cuadros

1. Introducción

Identificar grupos de proteínas que trabajan juntos, ó también llamados *complejos de proteínas* (PC), es importante porque permite comprender los mecanismos celulares con mucha precisión. Algunos mecanismos celulares *basados en proteínas* son responsables del transporte, comunicación, catálisis, almacenamiento, defensa y, entre otras, la reproducción de las células.

Muchas de estas actividades de nivel celular, son llevadas a cabo por la interacción física entre proteínas que forman complejos estequiométricamente estables. *Más del* 80 % *de las proteínas trabajan en complejos* [9] que pueden ser asociados a una tarea o función.

Estas dos últimas décadas, el crecimiento de las *bases de datos de proteínas y complejos* no se ha detenido. Los incontables beneficios que traen estos estudios alcanzan ciencias como la biología y la farmacología con la información necesaria para el desarrollo de medicamentos y tratamientos nuevos. Estas bases de datos de proteínas actualmente incluyen información sobre ontología genética, clasificaciones por función para proteínas aisladas y complejos, estructura cristalográfica y caracterización de secuencias.

Existen bases de datos para proteínas y PC seleccionados, que han sido ampliamente documentadas, verificadas experimentalmente y pueden ser consultadas como fuentes fiables y precisas, éstas son los *catálogos de proteína* o *Gold Standard Dataset* (GSD). Éstos catálogos o GSD existen en múltiples versiones y autores según año de publicación y especie biológica que describen, por ejemplo para levadura común existen: CYC2008 [3], Krogan core [4] ó Gavin [5].

A pesar de todos los avances actuales en esta materia, día a día siguen apareciendo nuevos organismos, nuevas proteínas y la posibilidad de identificar nuevos complejos. *La caracterización, aislamiento y verificación experimental de un complejo es un proceso costoso*, en términos de recursos económicos y de capital humano avanzado; por esos motivos, es importante desarrollar herramientas computacionales que permitan facilitar la predicción de complejos y ahorrar costos de experimentación, descartando computacionalmente (*in silico*), grupos de proteínas en los que se pueda demostrar baja o nula interacción.

Los avances en biotecnología han permitido el desarrollo de *muchas herramientas computacionales que buscan predecir interacciones físicas entre proteínas*. Algunas de estas herramientas están basados en grafos [1][13][14], otros en ontología genética, otras herramientas explotan la afinidad de propagación [18], otros predictores utilizan la similaridad de secuencia [19] y otros, utilizan clustering [15][16][17]. Para efectos de este trabajo, cualquiera de estas herramientas será un *predictor de complejos*.

Cada predictor de complejos tiene su propia base teórica, pero todos tienen la posibilidad de *comparar sus predicciones resultantes con GSD*. A partir de esta comparación los resultados se dividen en dos grupos, los verdaderos positivos

(TP) y falsos positivos (FP); el primer grupo son complejos reconocidos por la herramienta computacional y están registrados en los GSD. El segundo grupo de resultados no lo está.

El conjunto de los resultados FP es muy interesante de estudiar, ya que tienen características que los diferencian como complejos, según el predictor utilizado, y podrían potencialmente ser caracterizados como complejos y asociados alguna función.

Existen estudios [17] que enfatizan la importancia de la *caracterización de complejos a partir interacciones y la fisicoquímica molecular* [34], ya que es esencial para comprender, a alto nivel, la organización celular.

Para poder calcular y trabajar con características fisicoquímicas entre proteínas es importante conocer la *estructura espacial* de la proteína y su *secuencia de aminoácidos*, ésta información está disponible en forma libre en los bancos de proteínas. Algunos bancos son PDBe y RCSB Protein Data Bank.

Como la estructura de las proteínas es más estable y se conserva mejor que la secuencia durante la evolución [29] es más aconsejable *trabajar con las estructuras cristalográficas* (PDB) de cada proteína antes que las secuencias aisladas de datos espaciales.

Actualmente existen registrados $\sim$ 140,000 PDBs en los bancos de proteínas, cada semana se agregan $\sim$ 200 PDBs, cuando se separan las proteínas en estructuras primarias (cadenas) y dominios, las cifras aumentan a $\sim$ 330,000 y $\sim$ 500,000 respectivamente [29].

Al ser un formato tan rico en información, permite la construcción de metodologías y herramientas cada vez más exhaustivas desde una perspectiva biológica, por eso su producción tiene un crecimiento constante desde el año 2000; pero su uso computacional es muy costoso ya que trabajar con estructuras cristalográficas en la búsqueda de complejos de proteínas es un problema NP-Hard [51] [52], por ese motivo no es una técnica utilizada para la exploración de nuevos complejos de proteínas, pero este trabajo propone que es una buena aproximación para explorar los resultados FP de algún predictor.

1.1. Problema

Para estudiar proteínas y sus interacciones en la formación de PC se han desarrollado *gran cantidad de predictores computacionales*. Al comparar los resultados de un *predictor de complejos* con un *catálogo de proteínas*, o *GSD*, obtenemos un grupo interesante de resultados clasificados como complejos por el predictor pero que no están en los GSD, estos son los *resultados FP*.

Validar los resultados FP en forma aleatoria no es factible por el alto costo que tiene la investigación experimental de proteínas y complejos. Por ese motivo este trabajo propone la *construcción de un Ranking de resultados FP*, que según la *interacción espacial* entre proteínas y sus *características fisicoquímicas* ordene en primer lugar el resultado FP que tiene más posibilidades de ser complejo.

2. Objetivos

2.1. Objetivo Principal

El objetivo de esta memoria de título es diseñar e implementar un *Algoritmo de Ranking* para ordenar los resultados falsos positivos (FP) entregados por cualquier predictor de complejos, basado en una función que pondere las características fisicoquímicas de las proteínas que conforman cada resultado FP. Cuyo uso podría facilitar la planificación de la revisión experimental de los FP.

Objetivos Específicos

Para lograr cumplir con el objetivo principal se debe:

- Investigar técnicas, metodologías y herramientas para estudiar proteínas desde una perspectiva fisicoquímica.
- Diseñar e Implementar biblioteca que permita manipular y operar proteínas para construir el Ranking.
- Proponer e implementar un algoritmo de Ranking para resultados FP basado en información fisicoquímica.
- Aplicar y evaluar el algoritmo de Ranking sobre complejos de proteínas de Saccharomyces cerevisiae. Utilizando como método de predicción de complejos a DAPG [1] sobre una PPI network de Saccharomyces cerevisiae [12] (levadura común) y utilizando como GSD: CYC2008 [3].

3. Revisión Bibliográfica

Con el objetivo de encontrar la mejor estrategia para abordar el diseño de un *Algoritmo de Ranking* para resultados FP dados por algún predictor, se hace una revisión bibliográfica basada en tres ejes:

- **Marco teórico biológico:** tiene como objetivo crear una perspectiva desde el análisis general que entregan los predictores hasta la composición atómica que requiere este Ranking basado en un análisis exhaustivo.
- **Técnicas para detección de complejos:** tiene como objetivo conocer el estado del arte y los enfoques utilizados para la detección de complejos in silico.
- **Energía Libre de Gibbs:** tiene objetivo mostrar esta medida fisicoquímica como indicador de afinidad entre proteínas, así presentarla como función de orden para el Ranking.

3.1. Marco teórico biológico: desde el complejo hasta el carbono α

Los *predictores de complejos de proteínas* entregan como resultados FP una *lista de genes* que codifican proteínas que podrían interactuar juntas y formar complejos, basados en esto, se requiere conocer *cómo esta formado un complejo* y como descomponerlo hasta poder trabajar con *cada proteína individualmente* tal y como lo entregan los predictores.

Los PC interactúan con proteínas individuales u otros PC completando así *módulos funcionales*. Éstos incluyen funciones metabólicas, estructurales, inmunológicas, de transporte, regulación, señalización y movimiento.

Los PC se observan trabajando cohesionadamente, incluso en organismos muy simples como la *Saccharomyces cerevisiae* [10] (levadura común), que es un hongo unicelular. Es importante recordar que el código genético que rige la síntesis de las proteínas de todos los organismos es *universal*, tanto en seres vivos procariontes ó eucariontes. [11]. En términos prácticos un triplete (codón) del ADN codifica el mismo aminoácido en diferentes especies, sólo existen variaciones en los codones de inicio y término de síntesis al momento de transcribir y traducir desde ADN a Proteínas.

Las proteínas no se componen, en su mayoría, de una única cadena de aminoácidos, sino que se suelen agrupar varias *cadenas polipeptídicas* para formar proteínas multiméricas mayores. A esto se llama *complejos macro-moleculares* o PC.

Un PC, también llamado **estructura cuaternaria**, es una estructura compleja formada por varias cadenas agrupadas y cohesionadamente unidas entre ellas. Por ejemplo, una estructura cuaternaria es una fibra, formada por pro-

teínas más simples que tienen una distribución espacial, ó **estructura terciaria**, donde una de sus dimensiones es mayor a las otras dos, por eso es una fibra.

La *conformación espacial* de una proteína es la **estructura terciaria**, proteínas en este nivel de estructural, pueden ser fibrosas (e.g. colágeno y queratina) o globulares (no existe dimensión predominante, e.g ferrina o insulina).

La estructura terciaria, es específica para cada proteína y es lo que obtenemos al consultar *estructuras cristalográficas* (PDB). Es en este nivel estructural en donde se observan los puentes disulfuros creados por el azufre de la cisteína, que permiten un plegamiento tridimensional de un polímero lineal; este plegamiento se desarrolla, en parte espontáneamente, por la repulsión de componentes hidrófobos al contacto con el agua, la atracción de segmentos del polímero que están cargados, o la formación de puentes disulfuro; y por otra parte, ayudado por otras proteínas. Los polímeros lineales son la **estructura secundaria** de la proteína, y pueden ser clasificados en: hélice α y hoja β.

Todas las estructuras secundarias poseen una misma estructura química central, que consiste en una cadena lineal de aminoácidos (AA) o **estructura primaria**. Lo que hace distinta a una proteína de otra es la secuencia, de los 20 tipos de AA, que forman su estructura primaria.

Un aminoácido, como su nombre lo dice, son moléculas que tienen un grupo amino ($-NH_2$) y otro carboxilo o ácido ($-COOH$), que se unen por un carbono central ($-C_\alpha-$), las otras dos valencias del carbono central quedan saturadas con un átomo de hidrógeno ($-H$) y un grupo variable que llamamos radical ($-R$).

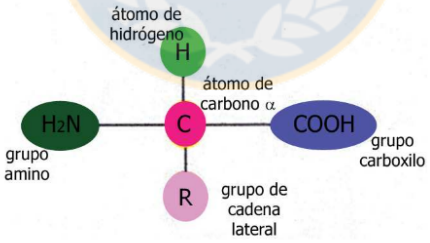

Figura 1: Fórmula general de un aminoácido

Espacialmente el carbono α presenta una configuración tetraédrica en la que éste se dispone en el centro y los cuatro elementos que se unen a él ocupan los vértices.

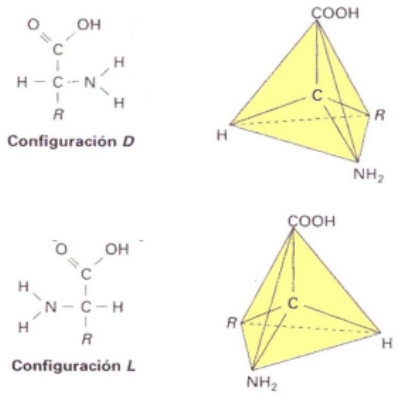

Figura 2: Disposición tridimensional de un aminoácido.

Un aminoácido es la unidad estructural básica de una proteína, conocer la disposición del carbono α nos permite conocer cual es la forma en que pueden interactuar dos proteínas diferentes dadas por un resultado FP, en donde podemos calcular la distancia entre cadenas alineando estos átomos carbono, y además nos permite calcular cual es el efecto que los radicales $(-R)$ que están próximos en el espacio, permitiendo así determinar las mejores combinaciones espaciales entre proteínas y por extensión caracterizar mejor los complejos.

3.2. Técnicas para detección de complejos: estado del arte

Actualmente existen muchos enfoques que buscan predecir la interacción entre proteínas, uno de los más utilizados es *buscar zonas de afinidad en la superficie* de una proteína donde puede producir enlaces fuertes al acercarse a otra (*protein docking*); basados en la cantidad de zonas de contacto y la fuerza de los enlaces que se producen, se aplican una serie de puntuaciones que favorecen o penalizan la inclusión de la proteína en el complejo que se esta caracterizando [22].

Hay trabajos [24] se proponen modelos para predecir todo el interactome (todos las interacciones de proteínas en una célula) a partir de operaciones basadas en docking.

Existen muchas herramientas disponibles para trabajar con docking que tienen un rendimiento aceptable en términos de tiempo y uso de memoria [25] [26], desarrolladas para estudiar la interacción de pequeños ligandos; por lo cual sería posible aplicar modelos de construcción incremental como el propuesto en [27] para buscar complejos, pero se presenta el *problema de la purificación de ligandos y proteínas objetivos.*

Las herramientas para computar docking de proteínas utilizando estructuras cristalográficas tiene la desventaja de requerir un *trabajo manual de purificación*, que consiste en identificar y aislar previamente el ligando y la molécula objetivo (target)[25].

Esta es una complicación en la tarea de exploración de complejos por *tres razones*. *Primero*, no se conoce con certeza qué proteínas conforman en complejo que se caracteriza como tal; *segundo*, no se conoce qué proteína actúa como objetivo y cuál como ligando; y si se pudiesen resolver los dos problemas anteriores, un *tercero* es que no se conoce el orden de agregación con el que se debe construir el complejo (árbol filogenético). Así la combinatoria crece complicando mucho más el problema que por si sólo ya es costoso de resolver en términos de memoria y tiempo.

Si bien existen *metodologías automáticas* basadas en docking para buscar complejos de proteínas como [22] sus operaciones se realizan sobre *conjuntos de proteínas y ligandos pre-definidos*.

Una alternativa a este problema sería utilizar *blind docking*, un conjunto de técnicas que busca iterativamente las mejores zonas de contacto, lamentablemente *no tienen buenos resultados en términos de exactitud* [28], como la exploración de complejos es una tarea que requiere la aplicación de la operación de docking reiteradamente, la propagación de errores es alta y los resultados poco confiables.

Existen propuestas basadas en el minado de grafos llamados Protein-Protein Interaction Networks (PPI networks), como el predictor utilizado para esta memoria: DAPS [1], que utiliza la búsqueda de subgrafos densos para buscar PC. Otras propuestas basadas en grafos [13] [14] tienen buenos resultados pero tienen problemas identificando complejos de formas irregulares. Otra desventaja de las herramientas basadas en grafos es que consideran afinidades generales ponderadas en un indicador de peso entre nodos, pero no realizan un análisis exhaustivo para la combinación de todas las proteínas de que forman un PC, esa pérdida de información los convierte en herramientas de baja resolución.

Otro enfoque para la detección de PC es el clustering aplicado a PPI networks, existen trabajos que explotan la afinidad de propagación [18], otros la ontología genética para agrupar proteínas [19], y otros [15] [16] [17] crearon un minado automático de PPI network basado en estrategias no especializadas de clustering. Al igual que las otras técnicas basadas en PPI network, son técnicas de baja resolución ya que no toman en cuenta la conformación natural de cada proteína y como interactúan entre ellas para determinar si son complejos o no.

Otro enfoque para abordar este problema es la *alineación estructural entre proteínas*, basado en que la estructura se conserva más que la secuencia durante la evolución, el alineamiento estructural múltiple puede ser más informativo que el alineamiento de secuencias (multiple sequence alignment) [7][29][30][31].

Para alinear computacionalmente proteínas se deben alinear geométricamente los C_α (*ver sección 3.1*) de una cadena de AA con los C_α de otras cadenas

(esta misma operación puede realizarse con los residuos (R) u otro componente de interés en el AA).

El alineamiento estructural busca la mejor combinación espacial de proteínas, donde maximizan o minimizan alguna medida tridimensional; las métricas físicas a este nivel más usadas son: root-mean-square deviation of atomic positions (RMSD) [32] que representa la distancia promedio entre los átomos de dos proteínas (regularmente su cadena principal, backbone atoms). La segunda medida global distance test (GDT) [33], trata la similaridad global entre proteínas que tienen zonas idénticas en sus cadenas pero diferentes estructuras terciarias (distribución espacial). La tercera métrica es TM-score [7], es una razón de similaridad que está en función del largo de los segmentos alineados.

La alineación de proteínas es ampliamente empleada en los estudios de biología estructural en tareas como modelado de proteínas, clasificación basadas en estructuras, entre otras [45][46][47].

En particular, existe una herramienta para alineamiento múltiple, mTM-align [29] que está construida basándose en TM-Align [7]. Esta herramienta al trabajar con estructuras cristalográficas permite fortalecer el valor de interacción cuando los residuos están cerca, y los debilita cuando están lejos, este factor es el Levitt–Gerstein (LG). Experimentalmente TM-score es muy robusto, el resultado normalizado que se obtiene al relacionar pares de proteínas aleatorias, no varía en función del tamaño de los polipéptidos comparados.

Para tratar el problema de Ranking de resultados FP, que aborda esta memoria, se utiliza la última técnica revisada: el *alineamiento estructural múltiple* (MStrPA) utilizando el en carbono α como parámetro de alineación. A continuación (*sección 3.3*) se explica como se puntuará la afinidad entre proteínas alineadas utilizando Energía Libre de Gibbs.

3.3. Energía libre de Gibbs: afinidad entre proteínas

Para comprender el reconocimiento molecular entre proteínas es necesario conocer los mecanismos fisicoquímicos que subyacen a la interacción entre ellas. En el caso sencillo en que dos proteínas P_1, P_2 , con mutua afinidad, que son mezcladas en una solución. Se puede formular la asociación dependiente de tiempo:

$$P_1 + P_2 \underset{K_{on}}{\overset{K_{off}}{\leftrightarrows}} P_1P_2 \tag{1}$$

Donde P_1P_2 representa el complejo formado por ambas moléculas, además K_{on} y K_{off} son las constantes de velocidad cinética de las reacciones de unión y disociación respectivamente. La medida para estas constantes son $M^{-1}s^{-1}$ o s^{-1}. En cuanto al equilibrio, la reacción de unión debe igualarse a la reacción de disociación (2).

$$K_{on}[P_1][P_2] = K_{off}[P_1P_2] \tag{2}$$

Donde $[P_i]$ es la concentración de equilibrio para la cualquier especie molecular P_i. Así la constante de disociación entre moléculas es:

$$K_d = \frac{K_{off}}{K_{on}} = \frac{[P_1][P_2]}{[P_1P_2]} = \frac{1}{K_b} \tag{3}$$

La unidad de medida de la constante de disociación K_d es mol (M), una baja constante de disociación esta acompañada de una reacción rápida, es decir alta afinidad de contacto entre proteínas.

Un sistema termodinámico de proteínas está compuesto por solutos (moléculas) y un solvente (agua o medio iónico). En este sistema hay interacciones complejas entre proteínas e intercambio de calor entre las sustancias (solutos, agua y íones amortiguadores), esta transferencia de calor está relacionada a los cambios de energía y se rigen por las leyes de la termodinámica. La energía libre de Gibbs (ΔG) es una medida termodinámica útil para caracterizar las fuerzas motrices, es un potencial que mide la capacidad de un sistema de realizar un trabajo máximo o reversible, en un medio isotérmico e isobárico , es una de las magnitudes termodinámicas más importante para la caracterización de fuerzas motrices[34].

En analogía con cualquier proceso espontáneo, la unión entre proteínas sólo se produce cuando el cambio de energía libre de Gibbs en el sistema es negativo al momento de equilibrio, a temperatura y presión constantes.

La energía libre estándar ΔG° está medida en condiciones de presión de 1 atm, una temperatura de 298 K y las concentraciones de reactivo efectivo de 1 M, esta se relaciona con la constante de disociación Kd de forma:

$$\Delta G^\circ = RTln(K_d) \tag{4}$$

Donde R es la constante universal de gases ideales ($1{,}987 calK^{-1}mol^{-1}$) y T es temperatura en Kelvin. A partir de la relación (4) se sabe que entre más baja la constante de disociación, más negativa es la energía libre de Gibbs estándar. Ésto indica que las constantes cinéticas (K_{on} y K_{off}) y su relación K_d determinan las propiedades termodinámicas del complejo, en otras palabras, la estabilidad del complejo y la afinidad entre proteínas [35][36]. Finalmente, a partir de un grupo de proteínas de un resultado FP, podemos alinearlas utilizando como parámetro sus carbonos α, y calcular la energía libre de Gibbs (ΔG°) como función de orden para el Ranking.

4. Algoritmo de Ranking

En esta sección se presenta el Algoritmo del Algoritmo de Ranking (RA), las estructuras que permiten las operaciones y el análisis del algoritmo considerando módulos que lo conforman, complejidades y características. El RA está formado por tres módulos principales:

- Mapeo GEN-PDB.
- Alineamiento de estructuras.
- Cálculo de Energía Libre de Gibbs.

La primera, mapea desde cada gen a una estructura cristalográfica asociada y reescribe los resultados FP que originalmente estan en términos de genes a PDB. La segunda, alinea múltiples estructuras cristalográficas que forman un resultado FP. La tercera, y última, calcula la energía libre de Gibbs en cada macromolécula alineada con éxito en la segunda etapa.

En el RA recibe de tres input:

- A: Lista de resultados FP.
- B: Criterio de búsqueda PDB (simple o max).
- C: Dirección para almacenar alineaciones.

Para encontrar las estructuras cristalográficas de cada gen dado en un resultado FP, se puede utilizar un criterio de búsqueda simple ó complejo. Es decir, el simple busca el PDB más aislado que se encuentre en la lista de referencias, el complejo es un PDB que está asociado a gen pero que puede contener otras proteínas. Para buscar y descargar la información se utilizó:

- UNIPROT (www.uniprot.org): Para obtener el número de ID. que relaciona genes y PDBs.
- RCSB Tools (https://rcsb.org/): Para descargar los PDB relacionados a cada gen.

Algoritmo de Ranking: Mapeo GEN-PDB

```
1: groups ← resultados FP (genes) y número de resultado FP
2: pdbs ← PDB de referencia y clave UNIPROT
3: genesid ← GEN de referencia y clave UNIPROT
4: unicos ← GEN de groups.
5: FOR(t: unicos)
6:     gid = Busca clave UNIPROT de t en genes
7:     IF(genes.end() == gid)
8:         tmpkey ← "Gen no tiene pdb"
9:     ELSE
10:        tmpkey ← almacena clave de referencia del gen buscado
11:    IF(tmpkey! ="Gen no tiene pdb")
12:        pid = Busca tmpkey en pdbs.
13:        IF(pdbs.end()==pid) Agrega a registro de genes sin PDB.
14:        ELSE agrega registro PDB a procesables.
15: END FOR
16: FOR(r: procesables)
17:      Descarga(r)
18: END FOR
19: pdbsupport←0;
20: FOR(s: groups)
21:    tkn ← Separa en palabras cada s.
22:    support ← tkn.size()
23:    FOR(d: tkn)
24:        aux ← cada pdb desde procesable con clave d
25:        pdbsupport++
26:    END FOR
27:    IF(pdbsupport/support ≥ 0.6)
28:        groupsbypdb ← agrega aux al contenedor
29:        pdbsupport ← 0
30: END FOR
31:
32: FOR(e: groupsbypdb)
33:     filter ← agrega cada palabra en e
34:     IF(filter.size()≥2)
35:        Crea input de alineación.
36:        Agrega filter al input de alineación.
37:        Limpia filter
38: END FOR
39:
40:
```

Para su construcción de este algoritmo se implementó una biblioteca en C++ con funciones para alinear proteínas [29]; otra función para calcular la energía libre de Gibbs [8], más algunas funciones auxiliares necesarias para construir el Ranking.

Algoritmo de Ranking: Alineación

41:
42: FOR(u: input de alineación)
43: Para cada u se genera una alineación
44: en la dirección C pedida como input
45: END FOR
46:

Algoritmo de Ranking: Energía de Gibbs

47: FOR(w: alineaciones)
48: Se calcula la energía de cada w, si el algoritmo encuentra zonas de contacto en la alineación registra en energyCollection.txt

49: END FOR
50: SORT(energyCollection.txt)

En el siguiente cuadro se detallan los contenedores utilizados, tipo y su función

Nombre	Tipo	Función
groups	unordered map	Almacena los resultados FP.
pdbs	unordered map	Contiene los pdb de la referencia.
genesid	unordered map	Contiene los genes de la referencia.
groupsbypdb	unordered map	Contiene los resultados FP en pdb.
procesables	unordered set	Contiene los pdbs que están disponibles para descarga.
unicos	unordered set	auxiliar para genes.
filter	unordered set	auxiliar para pdbs.

Cuadro 1: Contenedores utilizados en el Algoritmo de Ranking.

4.1. Complejidad de Algoritmo

El Algoritmo de Ranking esta separado en tres módulos funcionales, estos módulos funcionan en serie y son parte de un mismo programa, se presenta la complejidad en módulos ya que el módulo de alineación de estructuras múltiples es un problema que es NP-hard [51][52][53] y utiliza un algoritmo heurístico que no converge en términos del tamaño de entrada, como si lo hacen los otros dos módulos.

Módulo	Complejidad Temporal
Mapeo PDB-GEN	O(Nlog(N))
Alineación	Heuristico - No determinista
Energía Libre de Gibbs	O(Nlog(N))

Cuadro 2: Complejidad temporal para los módulos funcionales del algoritmo de ranking.

4.2. Diseño de Experimentos.

Para los experimentos fueron utilizados:

- Predictor de complejos DAPG[1].
- PPI network de Saccharomyces cerevisiae [12].
- Gold Standar CYC2008. [3]

En el siguiente diagrama de flujo se muestra la secuencia de tareas que realiza para la etapa de experimentación al que fue sometido el algoritmo de Ranking y una posterior etapa de validación.

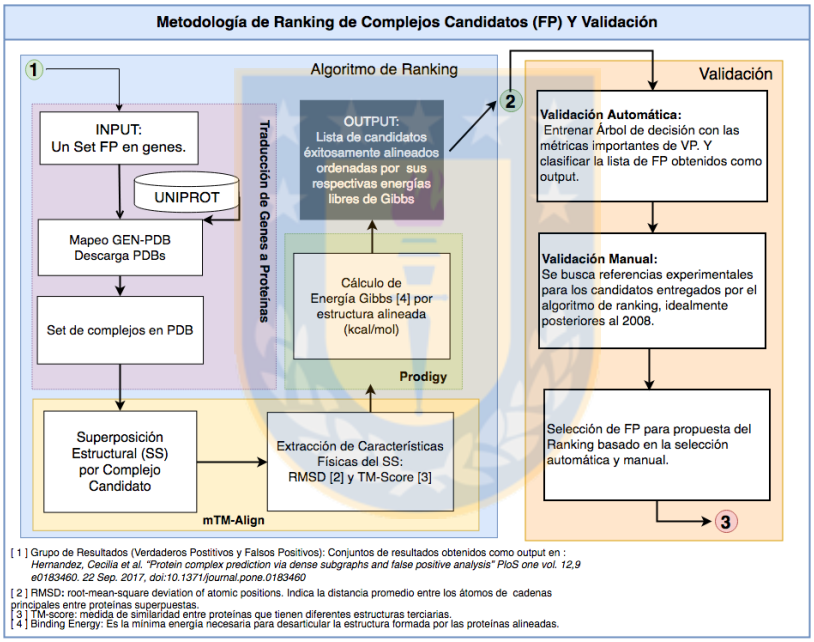

Figura 3: Flujo de trabajo del experimento diseñado para probar el algoritmo de ranking

Para la validación se usa un enfoque automático y uno manual. Para el enfoque automático se utiliza un árbol de decisión (DT) entrenado con las métricas importantes (sección 4.1) de todas las alineaciones de los complejos verdaderos positivos, que luego es utilizado para clasificar a los FP alineados exitosamente.

Para la validación manual se realiza una búsqueda de referencias experimentales de los candidatos FP en:

- La base de datos genómica de Saccharomyces de la Universidad de Standford [42].
- RCSB Protein Data Bank.[43]

4.3. Métricas Importantes.

Para realizar el algoritmo de ranking se utilizaron dos funciones importantes, la primera basada en alineación de proteínas [29] y la segunda en el cálculo de ΔG [8].

La función de alineación utiliza el TM-score y el RMSD. Qué están definidos como:

$$TMscore = MAX\left[\frac{1}{L_{target}} - \sum_{i}^{L_{ali}} \frac{1}{1+\frac{d_i}{d_0(L_{target})}}\right] \tag{5}$$

Donde:

- L_{min} es el largo de la cadena alineada más corta.
- L_{target} es el largo de la cadena en la que alineó.
- L_{ali} es el número de residuos alineados.
- d_i es la distancia entre el i-ésimo residuo. alineado.
- $d_0(L_{ali}) = 1,24\sqrt[3]{L_{ali}-15} - 1{,}8$

El RMSD es la distancia promedio entre los carbonos α alineados, tiene la desventaja de variar más ante cambios en la topología global, y no tanto en las variaciones locales, por lo cual tiene grandes propagaciones de error [32], pero complementado con el TM-score es un buen descriptor físico de la alineación. Se calcula como:

$$RMSD(u,v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|u_i - v_i\|^2} \tag{6}$$

mMT-Align[29], disponible en $C++$ [48].

La segunda función calcula una aproximación de la energía libre de Gibbs estándar a partir de un PDB con múltiples cadenas alineadas, ya que la afinidad de complejos proteína - proteína puede ser medida en términos de energía libre ($kcal/mol$)[39][36] que se definió en (*sección 3.3*).

La función que calcula $\Delta G°$ esta basada en una herramienta busca las áreas accesibles de la superficie soluble de la proteína utilizando NACCESS [40].

Adicionalmente se reporta que algunas interfaces de la proteína no interactúan (NIS). En donde la energía calculada entre superficies (ΔG°) es una simple regresión:

$$\Delta G^{\circ}_{predicted} = -0{,}09459 ICs_{charged/charged} - 0{,}10007 ICs_{charged/apolar}$$
$$+0{,}19577 ICs_{polar/polar} - 0{,}22671 ICs_{polar/apolar}$$
$$+0{,}18681\,\%NIS_{apolar} + 0{,}3810\,\%NIS_{charged} - 15{,}9433$$

El número $ICs_{xxx/yyy}$ corresponde a la cantidad de superficies de contacto entre dos estructuras terciarias, clasificando la superficie de acuerdo a si los residuos de esta son polares, no polares, cargados o naturales. Por ejemplo, $ICs_{charged/polar}$ es el número de superficies de contacto que una estructura tiene por una parte residuos cargados y en la otra residuos polares. Esta herramienta tiene un límite en la resolución de 5,5Å.

El cuadro 1 muestra un resumen de las métricas importantes y su interpretación.

Métrica	Rango	Interpretación
TM-score	[0,1]	0 son proteínas aleatorias, 1 proteínas identicas.
RMSD	[0 , $+\infty$[	Distancia promedio entre C_α
ΔG	] $-\infty$, 0 [	Entre más negativa, más estable es la molécula.

Cuadro 3: Resumen de métricas importantes.

5. Resultados

En término de los objetivos propuestos inicialmente se logró encontrar la forma de realizar un análisis exhaustivo basado en características espaciales y fisicoquímicas de las proteínas a partir de sus estructuras cristalográficas (PDB). Se implementó una biblioteca en c++ que permitió realizar experimentos con datos reales. La implementación del RA no fue absolutamente exitosa, ya que parte del problema es HP-hard y el algoritmo heurístico utilizado es no determinista, por lo que no hubo convergencia a una solución para todos los resultados FP con los que se experimentó. Al aplicar el RA la disponibilidad de resultados FP entregados por DAPS disminuyó considerablemente por dos razones: la primera tiene que ver con la disponibilidad de PDBs para cada gen buscado, y la segunda, por la convergencia del módulo heurístico en un periodo inferior a dos horas.

La primera disminución importante se realiza luego de mapear los genes con sus respectivos PDB, para efectos de este estudio se consideró un grupo válido aquel que tenía más del 60 % de los genes con PDB disponible.

El segundo filtro viene dado por la existencia de al menos dos PDB, se eliminan todos los candidatos que no puedan ser alineados. En este grupo en particular se detectó que hay convergencia de varios genes a un pdb, estos resultados no se exploraron ya que no podían ser alineados y evaluados en términos de energía libre de Gibbs.

El tercer filtro eliminó a los resultados FP que no terminaron la alineación en un tiempo igual a 2 horas.

El cuarto filtro elimina a las alineaciones que no exhibían zonas de contacto hidrosolubles para el cálculo de energía. En la figura 14 se aprecia como disminuyó la disponibilidad de grupos a lo largo del estudio.

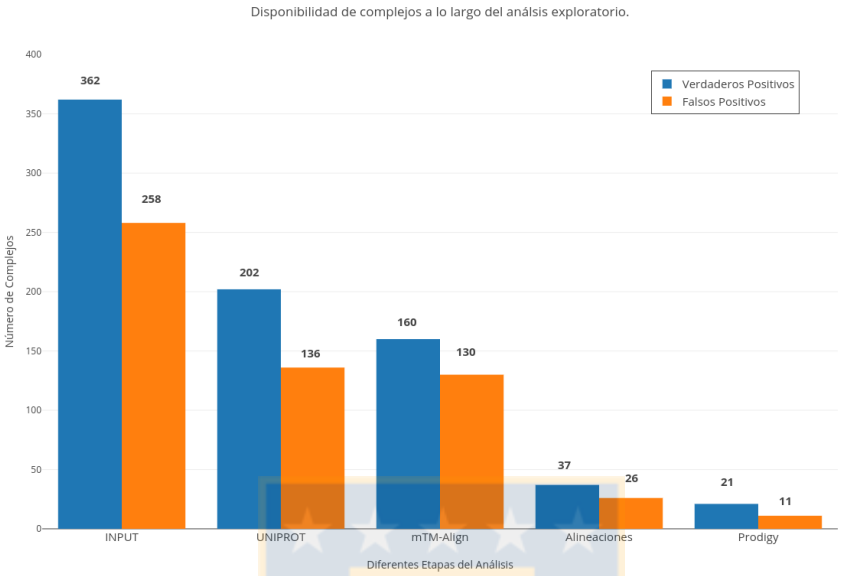

Figura 4: Disponibilidad de candidatos en las diferentes etapas del análisis.

De los 258 conjuntos FP sólo 11 complejos completaron el RA.

Para la etapa de Validación, al momento de entrenar el árbol de decisión, sólo se disponía de 21 complejos TP que terminaron el RA. Cómo resultado de la validación automática de FP, el clasificador indicó que todos los FP consultados eran complejos. La validez de estos resultados no son certeros ya que los árboles de decisión requieren muchos más datos de entrenamiento que 21 casos, ya que son son fácilmente ajustables. Podría utilizar otra metodología, como Random Forest, pero tener 21 casos de entrenamiento continua siendo insuficiente para técnicas automáticas.

Para comprender comprender mejor los resultados del clasificador se realizó una inspección de la distribución espacial generada al graficar: TM-score, RMSD y Energía para cada TP y FP. En la figura 15 se representó los casos que terminaron el RA sumados a los casos que terminaron la alineación pero no reportaron energía libre, debido a la falta de zonas hidrosolubles o de contacto (sólo cuentan con TM-score y RMSD).

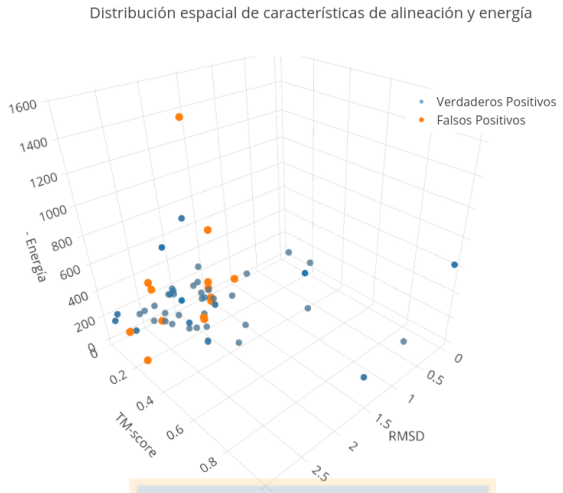

Figura 5: Distribución de características de alineación y energía para candidatos (FP) y complejos(TP).

Como se ve en la figura 15 los complejos (TP) muestran una distribución muy homogénea con respecto al TM-Score, esto puede interpretarse como que existen complejos de proteínas en donde sus proteínas sustrato pueden ser similares, aunque el 86 % de los TP tienen un TM-score entre 0 y 0.4, corresponde a proteínas no similares.

Para la distancia promedio entre átomos, tenemos complejos (TP) que están muy cerca ($0{,}0001x10^{-10}$ metros), aunque el 74,1 % esta entre $1{,}5x10^{-10}$ y $2{,}5x10^{-10}$ metros, el mismo segmento en donde se encuentra la mayoría de los FP alineados.

En términos de energía los complejos TP tienen una distribución concentrada entre los -60 $kcal/mol$ y los -300 $kcal/mol$, mismo segmento donde se encuentra el 80 % de los casos FP.

Viendo estos resultados se entiende porqué el árbol de decisión pudo clasificar a todos los FP como complejos. Son muy similares los conjuntos de entrenamiento y FP.

Por otro lado, durante la validación manual se encontraron referencias para 10 de los 11 candidatos seleccionados por el AR, en la siguiente sección se detallan los resultados obtenidos.

5.1. Ranking Generado para el análisis de FP de Saccharomyces cerevisiae

Este es el Ranking del los datos experimentales, usando como input los resultados FP de DAPG [1] para Saccharomyces cerevisiae usando los GSD CYC2008 [3]. Se presentan los candidatos ordenados por magnitud de energía (-E) decreciente. Sólo se incluyeron en el Ranking los resultados FP de los que se encontró documentación confirmando (DX) que son complejos, con año de publicación posterior al 2008, esto es importante ya que el catálogo de referencia utilizado es de ese año, y toda la documentación posterior a ese año confirma los resultados del Ranking.

Ranking	GEN_1	GEN_2	GEN_3	GEN_4	GEN_5	DX
1	MYO1	MYO2	MYO4	MLC1		6
2	MYO2	MYO4	MLC1	CMD1	MYO1	5
3	NUP2	KAP95	SRP1	KAP123		3
4	NUP2	KAP95	SRP1	NUP60		3
5	PSE1	KAP95	ULP1	SRP1		2
6	NHP2	CBF5	GAR1			5
7	NUP82	NSP1	NUP159			4
8	SHE2	SHE3	MLC1	MYO4		2
9	NRD1	AIR2	SEN1	NAB3		4
10	SHE2	SHE3	MLC1	CMD1	MYO4	3

Cuadro 4: Ranking resultante del análisis de complejos de Saccharomyces cerevisiae

Todas las referencias fueron obtenidas en:

- https://www.yeastgenome.org/
- http://www.yeastrc.org/
- https://www.rcsb.org/pdb/software/wsreport.do

Los dos primeros sitios especializados en el genoma de levadura común, el tercero es uno de los más importantes bancos de datos de proteínas.

6. Conclusiones

Se construyó un algoritmo de Ranking basado en dos operaciones con proteínas muy utilizadas en biotecnología: alineamiento de estructuras y cálculo de energía libre de Gibbs, pero no utilizadas frecuentemente para la detección de complejos, por ser herramientas muy exhaustivas y costosas.

La calidad de los resultados no fue la esperada debido a dos motivos: la disponibilidad de PDB no es alta para todos los genes que se buscó y el otro motivo es que uno de los módulos que compone el Ranking es heurístico y no logró convergencia de solución en un tiempo de ejecución igual a dos horas.

Este es un primer acercamiento a la alineación estructural, por lo que no se contaba con el excesivo tiempo de cómputo de las alineaciones, este algoritmo puede ser utilizado con otras especies biológicas, con otros GSD y otros tipos de predictores de complejos, basados en grafos o no. Que puedan complementarse con un análisis basado en estructuras cristalográficas alineadas y Ranking basado en energía libre de Gibbs.

El análisis automático de resultados resultó no ser discriminatorio, por el bajo número de casos para entrenamiento, se encontró suficientes referencias durante la validación manual que respaldan la idea de que esos grupos de proteínas seleccionados para el RA pueden interactuar como complejos.

El alineamiento estructural múltiple es una gran estrategia para abordar la factibilidad y la exploración de complejos, pero aún debe ser mejorado.

El tiempo de procesamiento para grupos pequeños de proteínas simples en promedio supera los 90 minutos, considerando que la Saccharomyces cerevisiae (levadura) no es un organismo complejo, el alineamiento no es una buena herramienta para estudiar características interactómicas de especies con muchas proteínas como el ser humano o el chimpancé. Lo que presenta una limitación importante en términos productivos para el análisis de gran cantidad de estructuras.

En el caso de mTM-align, escrito en c++, se puede explotar el paralelismo, pero debe cambiarse la implementación para computar la matriz de permutación, ya que como esta actualmente provocaría constantes condiciones de carrera.

Durante el análisis hubo un grupo de candidatos FP que tuvo disponibilidad de PDB sobre el 60 % pero no tenían más de 2 PDB para alinear, estos candidatos fueron eliminados en el primer filtro del proceso, pero tienen un alto potencial de ser complejos basados en que más del 60 % de los genes que tenían estaban agrupados en una sola estructura. Un análisis de esta naturaleza esta basado en la convergencia de genes a PDB, como no era el objetivo de esta memoria y se salía del análisis de características fisicoquímicas, ese grupo de proteínas no se estudió en profundidad.

Con respecto al cálculo de energía libre de Gibbs, la función se construyó sobre una herramienta cuya precisión no es muy alta, pero su rapidez compensaba

lo demoroso del alineamiento.

Existen muchos desafíos ad portas de la exploración a la biología celular mediante la informática. La información disponible tiene un volumen muy grande y es importante desarrollar metodologías de análisis basadas en algoritmos eficientes que permitan responder eficazmente a la demanda de procesamiento que actualmente requiere la biología computacional.

Sólo en el campo de los complejos de proteínas hay mucha información procesable, al ser un campo de estudio altamente colaborativo, hay mucho que revisar, leer y aprender.

Referencias

[1] Hernandez C, Mella C, Navarro G, Olivera-Nappa A, Araya J (2017) Protein complex prediction via dense subgraphs and false positive analysis. PLoS ONE 12(9): e0183460. https://doi.org/10.1371/journal.pone.0183460

[2] Collins, Sean R., et al. "Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae."Molecular & Cellular Proteomics 6.3 (2007): 439-450.

[3] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic acids research. 2009;37(3):825–831

[4] Krogan, Nevan J., et al. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae."Nature 440.7084 (2006): 637.

[5] Gavin, Anne-Claude, et al. "Functional organization of the yeast proteome by systematic analysis of protein complexes."Nature 415.6868 (2002): 141.

[6] Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., ... Liu, S. Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. International journal of molecular sciences, 17(2), 144. doi:10.3390/ijms17020144

[7] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. 2302–2309 Nucleic Acids Research, 2005, Vol. 33, No. 7 doi:10.1093/nar/gki524

[8] Xue L., Rodrigues J., Kastritis P., Bonvin A.M.J.J.*, Vangone A.*, "PRODIGY: a web-server for predicting the binding affinity in protein-protein complexes", Bioinformatics, doi:10.1093/bioinformatics/btw514 (2016).

[9] Berggård T, Linse S, James P. Methods for the detection and analysis of protein—protein interactions. Proteomics. 2007;7(16):2833–2842. doi: 10.1002/pmic.200700131

[10] Srihari, S. & Leong, H. W. A survey of computational methods for protein complex prediction from protein interaction networks. Journal of bioinformatics and computational biology 11, 1230002 (2013).

[11] Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. IUBMB Life. 2009;61(2):99-111.

[12] Collins, Sean R., et al. "Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae."Molecular Cellular Proteomics 6.3 (2007): 439-450.

[13] Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. & Kanaya, S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics 7, 207 (2006).

[14] Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, 471–488 (2006).

[15] Girvan, M. & Newman, M. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99, 7821–7826 (2002).

[16] Bader, G. & Hogue, C. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 1471–2105 (2003).

[17] Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes from protein-protein interaction networks. Nature Methods 9, 471–472 (2012).

[18] Frey, B. J. & Dueck, D. Clustering by passing messages between data points. Science 315, 972–976 (2007).

[19] Mukhopadhyay, A., Ray, S. & De, M. Detecting protein complexes in a ppi network: a gene ontology based multi-objective evolutionary approach. Mol Biosyst. 8, 3036–48 (2012).

[20] Runze Dong, Shuo Pan, Zhenling Peng, Yang Zhang and Jianyi Yang. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. W380–W386 Nucleic Acids Research, 2018, Vol. 46, Web Server issue. doi: 10,1093/nar/gky430.

[21] Vangone, A., & Bonvin, A. M. (2015). Contacts-based prediction of binding affinity in protein-protein complexes. eLife, 4, 454.doi : 10,7554/$eLife$,07454

[22] Comeau, Stephen R., et al. ÇlusPro: an automated docking and discrimination method for the prediction of protein complexes."Bioinformatics 20.1 (2004): 45-50.

[23] Smith, Graham R., and Michael JE Sternberg. "Prediction of protein–protein interactions by docking methods.Çurrent opinion in structural biology 12.1 (2002): 28-35.

[24] Vakser I. A. (2014). Protein-protein docking: from interaction to interactome. Biophysical journal, 107(8), 1785-1793.

[25] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Computational Chemistry 2009, 16: 2785-91.

[26] Brian Jiménez-García Jorge Roel-Touris Miguel Romero-Durana Miquel Vidal Daniel Jiménez-González Juan Fernández-Recio. LightDock: a new multi-scale approach to protein-protein docking.Bioinformatics. 2018 Jan 1;34(1):49-55. doi: 10.1093/bioinformatics/btx555.

[27] Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology, 261(3), 470-489.

[28] Ghersi, Dario, and Roberto Sanchez. Ïmproving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites."Proteins: Structure, Function, and Bioinformatics 74.2 (2009): 417-424.

[29] Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment.Bioinformatics. 2018 May 15;34(10):1719-1725. doi: 10.1093/bioinformatics/btx828.

[30] Layla Hirsh Damiano Piovesan Manuel Giollo Carlo Ferrari Silvio C. E. Tosatto. The Victor C++ library for protein representation and advanced manipulation. Bioinformatics, Volume 31, Issue 7, 1 April 2015, Pages 1138–1140, https://doi.org/10.1093/bioinformatics/btu773

[31] L. Martínez, R. Andreani, J. M. Martínez. Convergent algorithms for protein structural alignment. BMC Bioinformatics, 2007, 8:306.

[32] Shibuya T (2009). "Searching Protein 3-D Structures in Linear Time."Proc. 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2009), LNCS 5541:1–15

[33] Read, Randy J.; Chavali, Gayatri (2007). .ªssessment of CASP7 predictions in the high accuracy template-based modeling category". Proteins. 69 (S8): 27–37. doi:10.1002/prot.21662. PMID 17894351.

[34] Gibbs J.W. A method of geometrical representation of the thermodynamic properties of substances by means of surfaces. Trans. Conn. Acad. Arts Sci. 1873;2:382–404.

[35] Physicochemical bases for protein folding, dynamics, and protein-ligand binding. Li H, Xie Y, Liu C, Liu S Sci China Life Sci. 2014 Mar; 57(3):287-302.

[36] Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., ... Liu, S. Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. International journal of molecular sciences, 17(2), 144. doi:10.3390/ijms17020144

[37] Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins.J. Mol. Biol., 48, 443–453.

[38] Hubbard,T.J. (1999) RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins, 37, (Suppl. 3), 15–21.

[39] Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J. A structure-based benchmark for protein-protein binding affinity. Protein Sci. 2011 Mar;20(3):482-91. doi: 10.1002/pro.580. Epub 2011 Feb 16.

[40] Hubbard SJ, Thornton JM. NACCESS computer program. Department of Biochemistry and Molecular Biology. UK: University College of London; 1993. Available at: http://wwwbioinfmanchesteracuk/naccess/. Accessed on 2004. [Ref list]

[41] Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5?. Bioinformatics (Oxford, England), 26(7), 889-95.

[42] SGD, https://www.yeastgenome.org

[43] https://www.http://www.rcsb.org/

[44] https://github.com/bowbowbow/DecisionTree

[45] Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol.,247, 536–540.

[46] Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V.Proteins, 53, 334–339.

[47] Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. Science, 294, 93–96.

[48] $http://yanglab.nankai.edu.cn/mTM-align/$

[49] $https://www.uniprot.org$

[50] Anna Vangone, Alexandre MJJ Bonvin. Contacts-based prediction of binding affinity in protein – protein complexes. eLife 2015; 4 : $e07454doi$: 10,7554/$eLife$,07454

[51] Kolodny, R., Linial, N. (2004). Approximate protein structural alignment in polynomial time. Proceedings of the National Academy of Sciences, 101(33), 12201-12206.

[52] Inken Wohlers, Rumen Andonov, Gunnar W. Klau. Optimal DALI protein structure alignment. RR-7915, 2012, pp.20. ¡hal-00685824v1¿.

[53] Li S. C. (2013). The difficulty of protein structure alignment under the RMSD. Algorithms for molecular biology : AMB, 8(1), 1. doi:10.1186/1748-7188-8-1