Propuesta para resolver el problema de predicción de complejos de proteínas usando agentes inteligentes que aprenden usando técnicas de sistemas inmunológicos artificiales

Cristian S. Rocha
Laboratorio de Investigación y Desarrollo de Nuevas Tecnologías (LIDeNTec),
Gerencia de Sistemas y Telecomunicaciones (GSyT),
Administración Nacional de la Seguridad Social (ANSES)

5 de abril de 2009

Contexto

La ANSES, coordinada por la GSyT, busca explotar sus recursos computacionales no aprovechados fuera del horario laboral en benefício de otras instituciones que los requieran. En este contexto inició un convenio marco con el Hospital Italiano (HI) para desarrollar herramientas de bioinformática que puedan ejecutarse bajo éstas restricciones.

El LIDeNTec, dentro de este convenio, decidió aplicar técnicas de Inteligencia Artificial (IA) de fácil paralelización, como son los Agentes Inteligentes (AI), para resolver el problema de inferir estructuras de complejos de proteínas.

Resumen

En este trabajo se propone una técnica original del área de la inmunobiología para predecir las estructuras tridimensionales de complejos de proteínas - macromoléculas conformadas por dos o más proteínas encastradas para cumplir una función metabólica o estructual de la célula. La inmunobiología describe las técnicas usadas por el sistema inmune de los vertebrados para identificar patógenos. Esas mismas técnicas fueron recogidas por la IA para resolver problemas de clasificación. El trabajo recoge la capacidad innata del Sistema Inmunológico (SI) natural para aprender la geométría de las moléculas, pero no para identificar cuerpos extraños, sino con el objetivo de reconstruir complejos de proteínas comparando las geometrías aprendidas.

Palabras claves: reconstrucción de complejos, docking de proteínas, interacción proteína-proteína, agentes inteligentes, sistema inmunológico artificial.

1. Introducción

Muchos de los procesos biológicos requieren conocer la estructura tridimensional de complejos macromoleculares. Conocer la información estructural de dichos complejos puede ayudar a desarrollar nuevos compuestos farmacéuticos. Sin embargo los estudios cristalográficos y de Resonancia Magnética Nuclear (RMN) siguen una visión reduccionista donde las unidades de los complejos son más estudiadas que los mismos complejos. Esto se refleja en la base de datos de proteínas Protein Data Bank (PDB) donde existen una discrepancia en el número de unidades de proteínas (c. 3000) con respecto al número de complejos (c. 300).[7]

Es por ello que hay una fuerte necesidad de algoritmos de predicción de estructuras de complejos usando únicamente las coordenadas espaciales de los átomos de las proteínas que los componen. [7]

Existen diversas estrategias donde los algoritmos actuales se especializan: reconstruir complejos conocidos (*bounded docking*), y reconstruir complejos desconocidos (*unbounded docking*). En el primer caso existen técnicas que pueden recorrer todo el espacio de soluciones en un tiempo aceptable, mientras que la segunda, al requerir flexibilidad de los cuerpos involucrados, la cantidad de dimensiones y el tamaño del espacio de solución hacen al problema de muy difícil solución.[6]

En la biología de los vertebrados existe un sistema capaz de recorrer el espacio de soluciones en forma heurística y eficientemente: el Sistema Inmunológico (SI). Éste utiliza anticuerpos, proteínas generadas por células plasmáticas, con el objetivo de identificar sustancias antígenas. Cada anticuerpo tiene una única estructura que le permite unirse de forma específica a un único antígeno.[8]

La compleja heurística del SI fue simplificada y descrita por Dasgupta [3] para su uso en la computación como una técnica de Aprendizaje Automático (AA). De los diferentes niveles de complejidad detallados por Dasgupta [3], la teoría de redes inmunológicas se asociasa fácilmente con lo que se conoce como Agentes Inteligentes (AI) descripto por Fyfe and Jain [5].

En este trabajo se propone resolver el problema de predicción de estructuras de complejos usando redes de AI que aprenden, a través de un Sistema Inmune Artificial (SIA) , la superficie de la estructura de la proteína y su complementario. Toda la información recolectada se usa en un segundo paso para identificar las regiones de unión que permiten determinar el conjunto de complejos más probables de encontrar en la naturaleza.

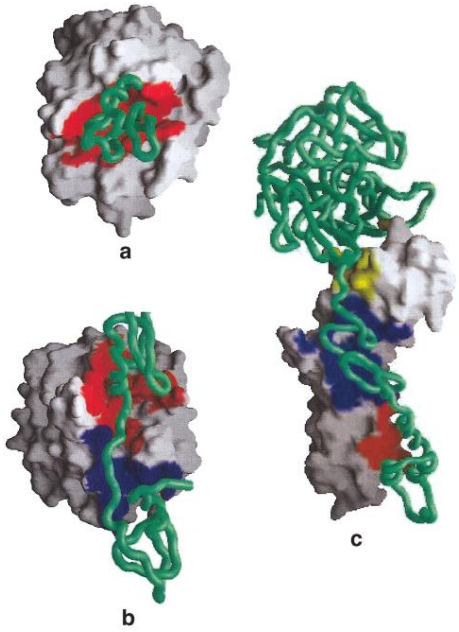

Figura 1: Parches identificados en complejos inhibidores de proteasas. Una de las poteínas del complejo tiene coloreada en la superfície los parches, la otra está representada por su backbone. Las proteínas listadas pueden encontrarse en PDB como A: 2pct, B: 1toc y C:1dan [2]

2. Lineas de Investigación y Desarrollo

La superfície geométrica que quiere reconocerse se la conoce como Superfície Accesible al Solvente (SAS). [9] Esta superfície es la frontera física entre las moléculas del solvente y la molécula en sí. Esta superfície no tiene una topológía única y puede variar según la flexibilidad de la molécula. Por ello es vital estudiar características, aunque sean parciales, de estas superfícies para poder identificar la interacción entre los cuerpos, y así predecir mejor la estructura de los complejos.

Chakrabarti and Janin [2] logro identificar parches de unión de aproximadamente $800\dot{A}^2$ (ver figura 1). Cada parche tiene un anillo de átomos accesibles al solvente y un núcleo de átomos no accesibles. Gracias a esta descripción podemos definir lo que para nuestro SIA es un anticuerpo y un antígeno.

Definición 1 *Un* anticuerpo *es una malla* $A_i \in \Re^3 \quad i \in [0, l]$ *capaz de envolver una region de la SAS de aproximadamente* $800A^2$ *con un cierto grado de tolerancia* t.

Definición 2 *Un* antígeno *es una superfície* $\bar{A}_i \in \Re^3 \quad i \in [0, n]$ *que representa la SAS de la proteína.*

Vease que la tolerancia t es el parámetro de flexibilidad de la afinidad a los antígenos. Cuando t tiende a cero, el anticuerpo es afín a un conjunto menor de proteínas. El valor de t, que consideramos biológicamente correcto es de $1{,}7\dot{A}$, el diámetro promedio de la molécula de agua.[9]

Para seleccionar nuestra lista de anticuerpos usamos el algoritmo Conalg (Clonación) [4]:

1. Generar una población aleatoria de anticuerpos
2. Presentación a los antígenos
   a) Evaluación de afinidad
   b) Selección de clones y expansión
   c) Maduración de la afinidad
   d) Reemplazar anticuerpos con otros generados
3. Ciclar el paso 2 hasta que se cumpla la guarda de parada.

Aunque nuestra SIA generá n anticuerpos aleatoreamente, requerimos de superfícies iniciales. Dos estrategias válidas para definirlas son:

- Si la muestra de aprendizaje es un conjunto de complejos conocidos, las superfícies iniciales son los parches descriptos por Chakrabarti and Janin [2]. Por ejemplo, si tomamos los complejos de la figura 1 se usarán los 6 parches identificados.
- Si la muestra de aprendizaje es un conjunto de proteínas se toma una parte de la superfície accesible al solvente.

A cada instancia de la nueva población se le incrementa un valor aleatorio $I_i \in \Re^3 \quad i \in [0, l]$ para generar la próxima generación de anticuerpos. Para que las instancias sean aprobadas hay que verificar que la superfície no exceda los $800\dot{A}$.

La evaluación de la afinidad de los anticuerpos usa la técnica de correlación para la predecir el docking [10]. Este algoritmo tiene la ventaja de ser exhaustivo y rápido para instancias pequeñas de datos. El resultado es una lista de configuraciones con su correspondiente puntuación dada por correlacionar el anticuerpo con el antígeno. La mejor puntuación es definida como la afinidad del anticuerpo al antígeno.

Del conjunto de anticuerpos generados y presentados, solo se eligirán m de mejor afinidad que se incluirán en el conjunto de anticuerpos ya elegidos. Aleatoreamente se eleminarán algunos y se quedarán con un conjunto reducido a m anticuerpos.

Es fácil de observar que este sistema de aprendizaje sesga su elección de anticuerpos a los cuerpos que les fueran presentados. Es por ello que se requiere de un sistema que flexibilice lo aprendido, distribuyendo el conocimiento entre varios generadores de anticuerpos. Estos generadores son AI.

Definición 3 *Los* Agente Inmunes *son AI que actuan como repositorios y generadores de anticuerpos.*

Estos agentes inmunes actúan individualmente para seleccionar los anticuerpos más afines para un determinado conjunto de antígenos, y actúan en conjunto para determinar probable interacción entre proteínas. En la figura 2 se presenta un diagrama de proceso de un agente donde se describe el aprendizaje y la comunicación con otros agentes.

Para identificar si dos anticuerpos son complementarios se usa la técnica de correlación descripta anteriormente.

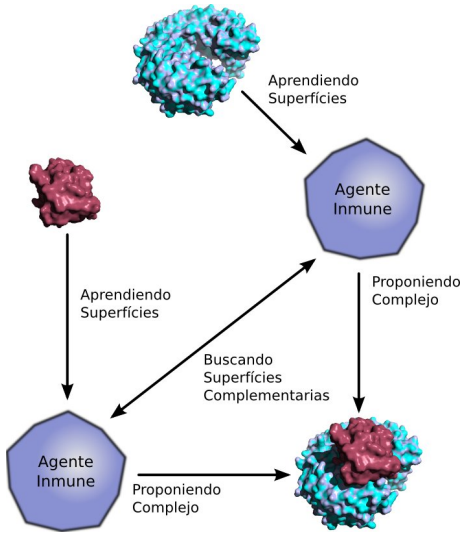

Figura 2: Diagrama de proceso de un agente inmune

3. Resultados Obtenidos/Esperados

Este trabajo se concentra en la utilización de técnicas de AA en el área de Bioinformática Estructural (BE). El desarrollo de estas técnicas están sujetas a no solo al estudio computacional de los algoritmos, sino también a pausibles interpretaciones biológicas, permitiendo un intercambio de conceptos e ideas entre ambas áreas. Al ser un modelo teórico preliminar solo se aproximó desde el punto de vista computacional. En una segunda etapa esperamos un contanste intercambio entre expertos de Microbiología, Bioquímica, Física y Computación.

El resultado obtenido es una reducción coherente del problema de predicción de estructuras de complejos al problema de identificación de antígenos que se describe en los SIA.

Esta aproximación deja abierta la discusión sobre la utilidad de técnicas de AA en el área de la BE, ya que en una exhaustiva búsqueda bibliográfica no se han encontrado una solución proveniente de esa área.

Se espera poder implementar los algoritmos a la brevedad y testear los resultados usando los datos descriptos por Andrusier et˜al. [1] como fuente de aprendizaje. Aunque la muestra está acotada a un conjunto reducido de complejos, sesgados a la facilidad de cristalización de las macromolécula, se espera conseguir un comportamiento aproximado al del Sistema Inmune Natural y así comparar resultados artificiales con resultados biológicos ya conocidos.

4. Formación de Recursos Humanos

El siguiente trabajo se desarrollará con la participación de los grupos de Bioinformática del Instituto de Ciencias Básicas y Medicina Experimental (ICBME) del Hospital Italiano (HI), el grupo de Bioinformática Estructural del Departamento de Química Biológica (DQB) y el grupos de Imágenes del Departamento de Computación (DC) ambos de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad de Buenos Aires (UBA), el grupo de Bioinformática del Centro de Estudio e Investigación (CEeI) de la Universidad Nacional de Quilmes (UNQ) y los grupos de Centro de Tecnología Médica (CTM) y Análisis Matemático de Imágenes (AMI) de la Universidad de Las Palmas de Gran Canarias (ULPGC), bajo la coordinación del Laboratorio de Investigación y Desarrollo de Nuevas Tecnologías (LIDeNTec).

Existe un gran interés de desarrollar la idea, prestar recurso y delinear aquellos experimentos que permitan un estudio exahustivo de esta técnica en casos reales. Es por ellos que cada grupo tomará una parte del trabajo y los desarrollará bajo un marco de tesis de grado.

5. Bibliografía

[1] Nelly Andrusier, Efrat Mashiach, Ruth Nussinov, and Haim J. Wolfson. Principles of flexible protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 73(2):271–289, 2008. doi: 10.

1002/prot.22170. URL http://dx.doi.org/10.1002/prot.22170.

[2] Pinak Chakrabarti and Joël Janin. Dissecting protein-protein recognition sites. *Proteins: Structure, Function, and Genetics*, 47(3):334–343, 2002. doi: 10.1002/prot.10085. URL http://dx.doi.org/10.1002/prot.10085.

[3] D. Dasgupta. An artificial immune system as a multi-agent decision support. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 4, pages 3816–3820 vol.4, 1998. doi: 10.1109/ICSMC.1998.726682. URL http://dx.doi.org/10.1109/ICSMC.1998.726682.

[4] Dipankar Dasgupta. *Artficial Immune Systems and Their Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1998. ISBN 3540643907. URL http://portal.acm.org/citation.cfm?id=552363.

[5] Colin Fyfe and Lakhmi Jain. Teams of intelligent agents which learn using artificial immune systems. *J. Netw. Comput. Appl.*, 29(2):147–159, 2006. ISSN 1084-8045. doi: 10.1016/j.jnca.2004.10.003. URL http://dx.doi.org/10.1016/j.jnca.2004.10.003.

[6] I. Halperin, B. Ma, H. Wolfson, and R.Ñussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, June 2002. ISSN 1097-0134. doi: 10.1002/prot.10115. URL http://dx.doi.org/10.1002/prot.10115.

[7] Thomas Lengauer, Raimund Mannhold, Hugo Kubinyi, and Hendrik Timmerman, editors. *Bioinformatics: From Genomes to Drugs*. Wiley-VCH, 1 edition, July 2001. ISBN 3527299882. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/3527299882.

[8] Kenneth Murphy. *IMMUNOBIOLOGY 7 PB (Janeway's Immunobiology) (Immunobiology: The Immune System (Janeway))*. Garland Science, November 2007. ISBN 0815341237. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/0815341237.

[9] A. Shrake and J. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–364, September 1973. ISSN 00222836. doi: 10.1016/0022-2836(73)90011-9. URL http://dx.doi.org/10.1016/0022-2836(73)90011-9.

[10] S. J. Wodak and R. Méndez. Prediction of protein-protein interactions: the capri experiment, its evaluation and implications. *Curr Opin Struct Biol*, 14(2):242–249, April 2004. ISSN 0959-440X. doi: 10.1016/j.sbi.2004.02.003. URL http://dx.doi.org/10.1016/j.sbi.2004.02.003.