Análisis de textos sobre interacciones entre proteínas[1]

Protein-Protein Interaction text analysis

Roxana Danger, Ferran Pla, Antonio Molina
Universidad Politécnica de Valencia
Camino de Vera, s/n
46022 Valencia
{rdanger, fpla, amolina}@dsic.upv.es

Resumen: Por su enorme interés en la genética, medicina y farmacología, detectar interacciones entre proteínas (PPI) es una de las áreas de investigación más importantes en el campo de las investigaciones biomédicas. De ahí que revista especial importancia el análisis (semi) automático de textos biomédicos que permita recuperar y mantener descripciones experimentales que justifican la presencia o ausencia tales interaccione. Tal es la finalidad del sistema on-line que se describe en el presente trabajo: dado un texto, se verifica su afinidad al tema sobre PPI, y varias entidades biomédicas son reconocidas y devueltas al usuario.
Palabras clave: extracción de información, reconocimiento de entidades.

Abstract: Protein-Protein Interaction (PPI) is one of the most important fields in biomedical research due to its enormous interest in genetics, medicine and pharmacology. Therefore, (semi)-automatic analysis of biomedical texts is critical for recovering and maintaining experimental descriptions which justify the presence or absence of such interactions. In this paper, an automatic, on-line system is described which, given a text, verifies if it corresponds to a PPI article, and recognizes various types of entities associated to this research context.
Keywords: information extraction, entity recognition.

1 Introducción

La interacción entre proteínas (PPI, por las siglas en inglés de Protein-Protein Interaction) es una de las ramas más importante en la investigación biomédica, pues permite predecir efectos secundarios a nivel celular, conocida la aparición de determinadas proteínas. La descripción de experimentos sobre PPI es el objeto del 5% de los artículos en PUBMED. Por ello es preciso contar con herramientas que permitan recuperar la información contenida en dichos textos y mantenerla en bases de datos que puedan ser fácilmente consultadas.

Las mejores soluciones a los principales problemas en esta área (clasificación de artículos sobre PPI, detección de genes y proteínas, normalización y recuperación de pares de proteínas que interactúan) han sido expuestas en las competiciones BIOCREATIVE (Hirschman et al., 2005), (Krallinger et al., 2009), (Leitner et al., 2009).

El presente trabajo describe el primer intento de creación de un sistema on-line para el reconocimiento de las entidades más importantes relacionados con PPI (*tipo de células, tejidos, interactores, roles biológicos, método de identificación de proteínas, método de detección de interacción y organismos*) y algunas relaciones entre éstas.

2 Descripción del sistema

El sistema que se describe a continuación, está relacionado con la clasificación de artículos sobre PPI, reconocimiento de proteínas y detección de pares de proteínas interactuando. Se han utilizado los conjuntos de test y entrenamiento provistos en BIOCREATIVE II.

[1] Este trabajo ha sido subvencionado por el proyecto TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) y por el programa “Juan de la Cierva” del Ministerio de Ciencia y Tecnología.

El módulo de clasificación de abstracts contiene un modelo SVM que considera características como 2 y 3-grams de palabras, co-ocurrencias de palabras en el conjunto positivo y negativo y mención de conceptos asociados a las interacciones moleculares. Los resultados experimentales arrojan un 0,74%, 0,63% y 0,68 para las medidas de precisión, recall y F-measure, respectivamente. Se ha incorporado el sistema BANNER (http://banner.sourceforge.net/) que permite la detección de genes y proteínas, cuyas prestaciones son de 88,66%, 84,32% y 86,43, respecto a la precisión, recall y F-measure. Los pares de proteínas interactuando son detectados a nivel de oración. Para ello, se verifica si existe un verbo, cuyo rol semántico sea el mismo de los comúnmente utilizados en PPI, que relacione las proteínas encontradas en una oración.

Adicionalmente, el sistema que describimos cuenta con un reconocedor de otras entidades relacionadas con PPI, basado en diccionario y matching no exacto. Las entidades reconocidas son: *tipo de células, tejidos, interactores, roles biológicos, método de identificación de proteínas, método de detección de interacción* y *organismos*. Los términos asociados a estos tipos de entidades, son indexadas por Lucene, y luego, haciendo búsquedas sobre dicho índice, las frases más cercanas en el texto a aquellos términos indexados son recuperadas. Varias consideraciones acerca de la similitud frase-término permiten finalmente aceptar una frase de un texto como término de interés en el contexto PPI.

La Figura 1 muestra la respuesta dada por el sistema para un párrafo asociado al contexto de PPI. Nótese que la última interacción no es correcta pues es descrita en un contexto donde hay negación.

Detectar este tipo de interacciones, entre otras más complejas (negadas, generalizadas, entre muchas proteínas, utilizando sinonimias, o cuando se empleen expresiones del habla con oraciones subordinadas y coordinadas), es objetivo del trabajo futuro.

Un módulo relacionado con la normalización de proteínas está en fase de desarrollo. Su inclusión garantiza la extracción de información de manera consistente, ofreciendo la posibilidad actualización de bases de datos de forma casi automática.

Así mismo, generalizar el proceso de extracción de información sobre interacciones entre proteínas a nivel artículo, no sólo a nivel de abstracts, permitirá recuperar una mayor cantidad de descripciones experimentales sobre PPI. Ello se traduce, en poner al alcance de los biomédicos información completa y actualizada, permitiéndoles, detectar nuevas interacciones moleculares de forma más rápida y fiable, repercutiendo ello, a su vez, en los campos de la medicina, la genética y la farmacología.

We screened proteins for interaction with <**protein> presenilin (PS) 1 </protein>**, and cloned the full-length cDNA of **<organism>human<organism> <protein> delta-catenin </protein>**, which encoded 1225 amino acids. **<interactDetectMethod> Yeast two-hybrid assay </interactDetectMethod>**, **<interactDetectMethod> GST binding assay </interactDetectMethod>** and **<interactDetectMethod> immunoprecipitation </interactDetectMethod>** demonstrated that **<protein> delta-catenin </protein>** interacted with a hydrophilic loop region in the endoproteolytic C-terminal fragment of **<protein> PS1 </protein>** , but *not* with that of **<protein> PS-2 </protein>**.

Tipos de entidades identificadas:	*Interacciones*:
1. Protein	1. delta-catenin - presenilin 1
2. Organism	2. delta-catenin - PS-1
3.Interaction Detect Method	3. delta-catenin - PS-2

Figura 1: Ejemplo de respuesta dada por el sistema.

Bibliografía

Hirschman L., Yeh A., Blaschke C., Valencia A. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology BMC Bioinformatics, 6(Suppl 1):S1.

Krallinger M., Morgan A., Smith L., Leitner F., Tanabe L., Wilbur J., Hirschman L., Valencia A.. 2008. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology*, 9(Suppl 2):S1.

Leitner F., Krallinger M., Valencia A. 2009. The BioCreative II.5 Online Challenge. http://www.biocreative.org/media/store/files/2009/The_BioCreative_II.5_Online_Challenge.pdf.