PPInterFinder – A Web Server for Mining Human Protein-Protein Interaction

Kalpana Raja, Suresh Subramani, Jeyakumar Natarajan
Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, India

Background of the system

PPInterFinder is a web server that integrates text mining and information extraction techniques to mine information on Protein-Protein Interaction (PPI) from biomedical literature. The server is specific to human PPI and is currently under development at the Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, India. The server is expected to be more accurate to extract human PPI from biomedical literature as it uses two background modules namely, NAGGNER (1) for Named Entity Recognition (NER) and ProNormz (2) for normalization. These two modules are highly specific to human proteins and have already been developed in the same lab. NAGGNER and ProNormz are available for biocuration purposes as two separate tools. PPInterFinder will be available from 1 March 2012.

The input for PPInterFinder can be a biomedical abstract(s) or full text, and the corresponding output is the mined PPI information. The protein names are being normalized to official symbol in order to offer unique reference. In addition to the extracted information, the server displays other related PPI information for every protein in the sentence together with Gene ID and PubMed ID for cross references. The server implements a co-occurrence based algorithm on the backend database to store the relevant PPI information found across biomedical literature. The system also identifies the presence of negation using word’s proximity in order to improve the accuracy of the system.
Example of extracted PPI information

All the listed articles discuss on human PPI on different proteins. PPIterFinder displays the extracted information with PubMed ID and provides a link to check other related interactions for every protein.

For example, PubMed ID: 11556834 provide information on physical interaction between two human proteins CDKN2A and CDK4. Likewise, PubMed ID: 12628926 discuss the association of TBL1XR1 and NCOR1 through two independent interactions. PPIterFinder uses Natural Language Processing (NLP) approach to recognize more appropriate interacting keyword (associates) when more than one interacting keywords are present in the same sentence (associates and interactions).

The backend database of PPIterFinder gets updated every time a new PPI is hit. Interaction of a protein for example, CDKN2A with all reported proteins can be retried through related interactions. PPIterFinder gives both GeneID and PubMedID for cross references. Co-occurrence indicates the number of times the same PPI information appeared in biomedical literatures. Such a database can even provide an advanced list of PPI information.
Additionally, PPInterFinder facilitates the display of complete abstract with the mined PPI information is highlighted. While the protein names in the abstract are normalized to official symbol, the system displays them when the mouse is moved over.

### List of related interactions for CDKN2A

<table>
<thead>
<tr>
<th>Protein A (Gene ID)</th>
<th>Protein B (Gene ID)</th>
<th>PubMed ID</th>
<th>Co-occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKN2A (1029)</td>
<td>CDK4 (1019)</td>
<td>11556834, 8805225</td>
<td>2</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>MDM2 (4191)</td>
<td>8805225</td>
<td>1</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>CDK6 (1021)</td>
<td>11556834</td>
<td>1</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>MYC (4609)</td>
<td>15361884, 10837489, 9111042</td>
<td>3</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>UBE2L3 (7332)</td>
<td>12082609</td>
<td>1</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>MAPK8 (5599)</td>
<td>16007099</td>
<td>1</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>PSMC3 (5702)</td>
<td>14665636, 15577913</td>
<td>2</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>HIP1A (3091)</td>
<td>11382768, 15674338</td>
<td>2</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>NCOA1 (9611)</td>
<td>15729358</td>
<td>1</td>
</tr>
<tr>
<td>CDKN2A (1029)</td>
<td>YY1 (7528)</td>
<td>15210108</td>
<td>1</td>
</tr>
</tbody>
</table>
System Adaptability and Interactivity

PPInterFinder will be available as a free web-server, integrating a tool based environment to mine the PPI information from a given biomedical literature and a database to store the mined information. The backend database of PPInterFinder differs from the normal PPI databases where the stored information are limited and rarely updated. However in PPInterFinder, the mined PPI information are updated to the database in two different ways based on their existence in the database; (i) New entry and (ii) Updated entry. Also the system increments the co-occurrence time and adds the corresponding PubMed ID for later references. Large numbers confirm more proven PPI information published across biomedical literatures.

PPInterFinder is under development using Java and MySQL. It is available for use on any commonly known computer platforms. The proposed system provides an option to screen multiple biomedical abstracts for batch annotation and PPI extraction per execution. PPInterFinder also provides link to PubMed database for abstracts and NCBI database for proteins. Though the original names are normalized to official symbol, the system displays the name as in abstract when the mouse is moved over. In addition, the system retrieves a list of all related PPI information to every protein along with GeneID and PubMed ID.

Performance

The internal modules of PPInterFinder namely, NAGNNER and ProNormz have been evaluated on two gold standard corpuses namely NLPBA and BioCreAtIvE for human proteins. Besides we tested their performance on our own corpus of 750 MEDLINE abstracts related to human proteins and protein
kinases. The performance of the modules is evident from their high precision, recall and f-score values (1, 2) for human proteins. We have created an initial training dataset (currently comprising of 6059 sentences with human PPI information) for PPIterFinder filtered from the PPI dataset of IntAct database. As the proposed system considers the presence of negations (e.g. not interacting, not associating) while deciding PPI information, the accuracy of the system is expected to be much better. The evaluation of the complete system will be published in Mar 2012.

**Proposed task for TRACKIII for PPIterFinder**

1-Given a set of abstracts describing PPI information for human and other proteins. These abstracts are not yet curated using the system (i) identify the human proteins appearing in the abstract (NER) and replace with official symbol (Normalization). These are called as curatable or positive set; and (ii) for curatable abstracts, extract PPI information.

The task will be run manually and using the PPIterFinder system (but not by the same curator).

**Manual task:** the user will be given the list of abstracts from Pubmed and a spreadsheet to list the proteins appearing in the abstracts along with the official symbol from Entrez Gene Database and the type of organism as human or others (Entrez Gene).

**Using eFIP:** curator would assert the curatable set of abstracts, and validate the extracted information.

**Input:** MEDLINE abstracts related to human or other PPI.

Below is a list of PMIDs for sample abstracts.

11556834  
12628926  
15494311  
15893728  
11382768  
8805225  
11556834  
15361884  
22172674  
22166218

**Output:**

1-Provide a tab delimited file along with the following information:

PMIDs curatable (contain at least two human proteins and an interacting keyword in the sentence describing PPI information) along with the following information:
PMID | Curatable | Protein 1 | Protein 2 | Other protein (s) | Interacting Keyword

If any information is not available in the given sentence, a dash should be used for that field.

An example of this file based on set of curatable articles from the list above is shown here.

<table>
<thead>
<tr>
<th>PMID</th>
<th>Curatable</th>
<th>Protein A</th>
<th>Protein B</th>
<th>Other Protein(s)</th>
<th>Interacting Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>11556834</td>
<td>yes</td>
<td>CDKN2A</td>
<td>CDK4</td>
<td>-</td>
<td>interaction</td>
</tr>
<tr>
<td>15494311</td>
<td>yes</td>
<td>CHUK</td>
<td>NCOR2</td>
<td>NFkB1</td>
<td>phosphorylate</td>
</tr>
<tr>
<td>12628926</td>
<td>yes</td>
<td>TBL1XR1</td>
<td>NCOR1</td>
<td>-</td>
<td>associates</td>
</tr>
</tbody>
</table>

References
