PPAXE FACILITATES HUMAN-CURATION OF PROTEIN-PROTEIN INTERACTIONS FILTERED FROM THE SCIENTIFIC LITERATURE

Castillo-Lara, Sergio; Abril, Josep F.

Computational Genomics Lab · Universitat de Barcelona · IBUB

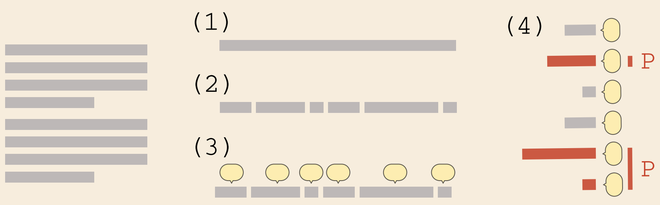
SUMMARY

Protein-protein interactions (PPIs) are crucial to build models for understanding many biological processes. Although several databases hold many of these interactions, exploring them, selecting those relevant for a given subject, and contextualizing them can be a difficult task for researchers. Extracting PPIs directly from the most recent scientific literature sources can be very helpful for providing such context, as the sentences describing these interactions may give insights to researchers in helpful ways.

We have developed a python module and a web application, PPaxe, that allows users to extract PPIs and protein occurrence from a given set of PubMed and PubMed Central articles, based on abstracts and full-texts respectively. PPaxe tokenizes and annotates the sentences with StanfordCoreNLP [1] and then distills a number of features that are analyzed by a Random Forest classifier. Finally, it presents the results of the analysis in different ways to help researchers export, filter and analyze the interactions easily.

PROTEIN NAME RECOGNITION

PPaxe uses the StanfordCoreNLP program suite to tokenize the sentences, annotate them using Part-of-Speech (POS) tags, and to recognize the proteins in the sentences. In order to do so, the conditional random field (CRF) provided by the package was trained over three datasets (AImed [2], MedTag [3] and BioInfer [4]), which contain annotated abstracts from scientific publications.



STANFORD CORE NLP

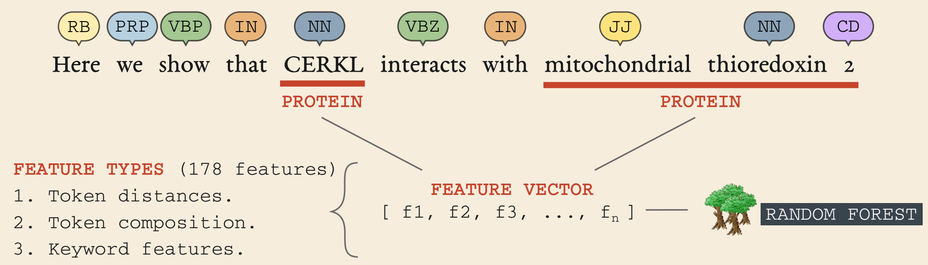
1- Sentence extraction.
2- Tokenization.
3- POS tagging.
4- Protein recognition.

The performance of the protein tagger was assessed using 2-fold cross-validation.

PRECISION	RECALL	F1
74.5%	70.0%	72.23%

INTERACTION RETRIEVAL

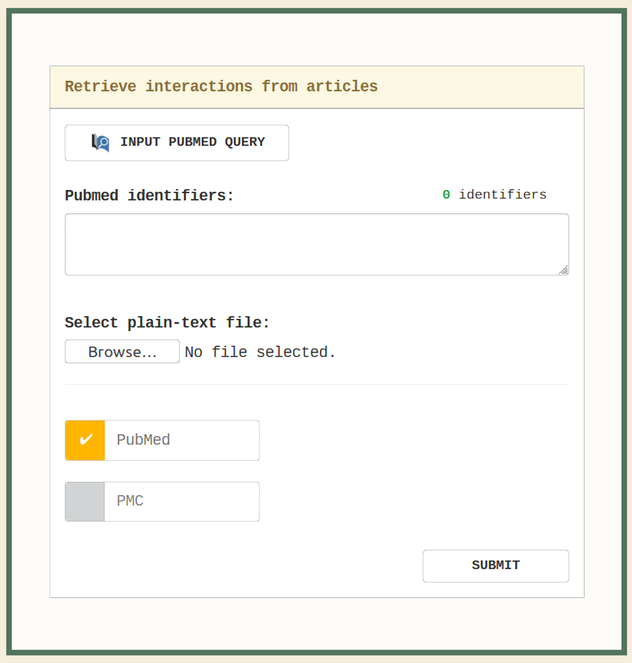
PPaxe selects those sentences containing co-ocurring proteins, and then computes several features. These features are based on distances between different sentence POS tokens, the composition of said POS tags between and around the proteins, and counts for different keywords. By basing our predictions on these basic features, PPaxe frees the StanfordCoreNLP of predicting the more computationally-intensive syntactical relationships. A random forest classifier was then trained over AImed, LLL-challenge [5], and the Bioinfer datasets, which contain annotated PPI.



The interaction extraction performance was assessed using 10-fold cross-validation:

PRECISION	RECALL	AUC
76.50%	58.68%	0.92

PPAXE APPLICATION



PPAXE WEB FORM

PPaxe is distributed both as a Python module and as a web application. The PPaxe web application can be downloaded and installed on any computer, thanks to the provided Docker containers. The web application consists of a form, where users can introduce PubMed identifiers from which to retrieve PPI. Users can look for interactions either on abstracts, on full-text articles (if available on PubMed Central) or on uploaded plain-text files.

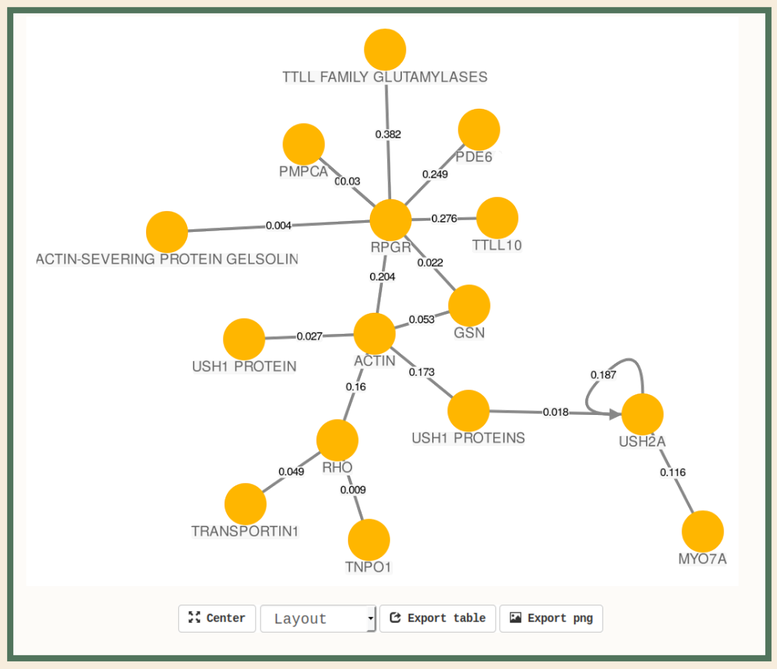
QUERY PUBMED FROM PPAXE

PPaxe allows researchers to define PubMed queries directly from the application, performing the analyses without leaving PPaxe. These queries support standard PubMed field tags, so that users can restrict their queries to organisms, authors, specific years, and more.

NETWORK VISUALIZATION

In the results page, PPaxe will also display an interactive graph panel of the retrieved PPIs; made using the JavaScript library cytoscape.js. Users can modify the layout of the displayed nodes (either manually or by using an option menu with a few layout presets). Finally they can export the results to several graphical formats.

The distilled network will contain the PPIs retrieved from the literature, with the confidence value of the predicted interaction shown on the interaction edges.



Confidence	Off.symbol (A)	Off.symbol (B)	PMid	Year	Sentence
0.649	CNNM4	IQCB1	29322253	2018	Here we show that **CNNM4 interacts** with **IQCB1** , which causes Leber congenital amaurosis (LCA) when mutated .
0.644	CEP78	FAM161A	27588451	2017	Interaction studies also showed that **CEP78 binds** to **FAM161A** , another ciliary protein associated with retinal degeneration .
0.418	PIH1D3	DNAAF2	28041644	2017	Further , **PIH1D3 interacts** and **co-precipitates** with cytoplasmic ODA/IDA assembly factors **DNAAF2** and DNAAF4 .
0.378	PIH1D3	DNAAF4	28041644	2017	Further , **PIH1D3 interacts** and **co-precipitates** with cytoplasmic ODA/IDA assembly factors DNAAF2 and **DNAAF4** .

RETRIEVED INTERACTIONS

Once the analysis is completed, PPaxe will redirect users to the results page, where the retrieved interactions will be accessible through a searchable table.

In this table, PPIs will be sorted by their confidence score, and researchers can access the sentence describing those PPIs, along with the source article, allowing them to gather and curate the results easily.

EXPORT OPTIONS

PPaxe results from the analyses can be retrieved in different formats. A PDF report containing all the information available through the results page (the interactions table, the proteins ocurrence, the graph visualization, and several other plots). Users can download a CSV table containing the retrieved interactions; and a static representation of the network in PNG format.

ACKNOWLEDGEMENTS

This work was supported by Spanish Ministry of Economy (BFU2014-56055P); Generalitat de Catalunya (2014SGR687, 2018SGR1455); Predoctoral fellowship by AGAUR (FI-FDR, 2017FI_B_00191).

REFERENCES

[1] Manning, C. et al. (2014). In A.C.L., 55–60.

[2] Bunescu, R. et al. (2005). Artif. Intell. Med., 33(2), 139–155.

[3] Smith, L. H. et al. (2005). In Proc. of the ACL-ISMB. Workshop on LinkingBiological Literature, 32–37.

[4] Pyysalo, S. et al. (2007). BMC Bioinformatics, 8 .(1), 50.

[5] Nédellec, C. (2005). In Proc. of the Learning Language in Logic 2005 Workshop at the Int.Conf. on Machine Learning.

AVAILABILITY

compgen.bio.ub.edu/ppaxe

PPAXE · docker · python Package Index