Internal Report

Deliverable D3.1a:

Refined vocabulary for cross-context knowledge discovery in plant immune signalling

National Institute of Biology Version 1 FINAL

Abstract: To improve vocabulary for cross-context knowledge discovery in plant immune signalling, we have identified several steps where the vocabulary should be improved. Three important steps were chosen, namely searching query to retrieve the articles from open access databases, top keywords in OntoGen and SVM keywords in Ontogen. For each separate step, we performed several cycles of refining, often including manually curated corrections.

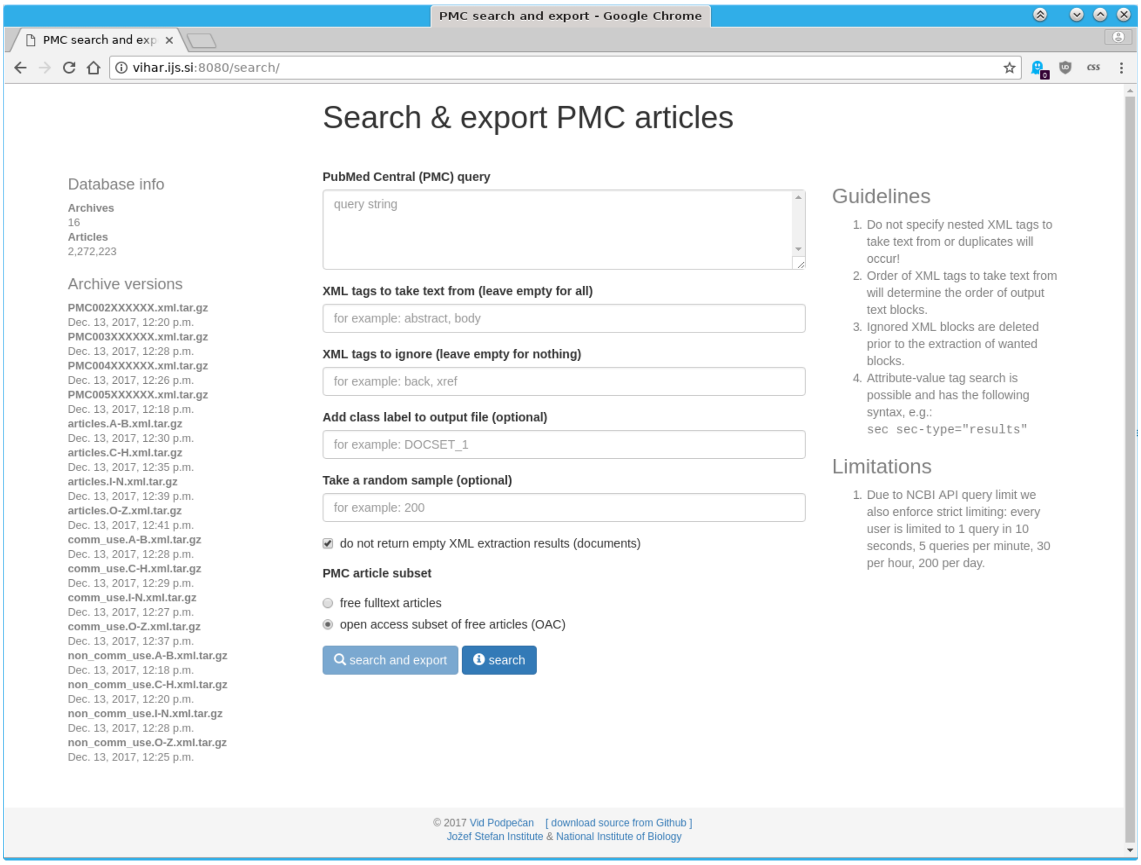
Document administrative information	
Project acronym:	HinLife
Project number:	J7-7303
Deliverable number:	D3.1a
Deliverable full title:	Refined vocabulary for cross-context knowledge discovery in plant immune signalling
Document identifier:	HinLife D3.1 v1.FINAL
Lead partner short name:	NIB
Report version:	Version 1, FINAL
Report preparation date:	12/02/2018
Lead author:	Maruša Pompe Novak, Vid Podpečan, Dragana Miljković, Senja Pollak, Bojan Cestnik
Co-authors:	Nada Lavrač
Status:	Final

Introduction

It is known that plants have the circadian rhythm, which means that some of their genes are differently expressed in the morning, at noon and in the evening. We have noticed that plants react differently to the infection with the pathogen if they are infected in different parts of the rhythm. We can conclude that genetic networks that are important for the defence of plants against pathogens are differently expressed in the different parts of the day.
We focused on knowledge discovery by combining knowledge from two different domains (plant defence and circadian rhythm) to get new insights and derive new conclusions. Our cross-domain literature mining methodology includes three complementary text mining tools: literature (text) preparation and extraction tool, clustering and topic ontology creation tool OntoGen and a cross-domain bridging terms exploration tool CrossBee. Using OntoGen we try to identify outlier documents while CrossBee allows bridging concepts exploration and identification.

Data preparation

We have developed a web application which allows for search, preprocessing, extraction and download of PubMed Central (PMC) free-text articled. PMC archives ~4.7 million articles of which more than 2 million have free full-text available for download. Moreover, the majority of the free full-text collection is released under the open access licence (OAC). NCBI provides an API which allows searching and downloading full texts of articles. However, API query rates are limited to allow uninterrupted access to NCBI website and web tools. Therefore, NCBI also makes their free full-text articles collection available as a set of large archives of compressed XML documents.


PMC search and export - Google Chrome
PMC search and exp ×
vihar.ijs.si:8080/search/

Search & export PMC articles

Database info

Archives
16
Articles
2,272,223

Archive versions

PMC002XXXXXX.xml.tar.gz	Dec. 13, 2017, 12:20 p.m.
PMC003XXXXXX.xml.tar.gz	Dec. 13, 2017, 12:28 p.m.
PMC004XXXXXX.xml.tar.gz	Dec. 13, 2017, 12:26 p.m.
PMC005XXXXXX.xml.tar.gz	Dec. 13, 2017, 12:18 p.m.
articles.A-B.xml.tar.gz	Dec. 13, 2017, 12:30 p.m.
articles.C-H.xml.tar.gz	Dec. 13, 2017, 12:35 p.m.
articles.I-N.xml.tar.gz	Dec. 13, 2017, 12:39 p.m.
articles.O-Z.xml.tar.gz	Dec. 13, 2017, 12:41 p.m.
comm_use.A-B.xml.tar.gz	Dec. 13, 2017, 12:28 p.m.
comm_use.C-H.xml.tar.gz	Dec. 13, 2017, 12:29 p.m.
comm_use.I-N.xml.tar.gz	Dec. 13, 2017, 12:27 p.m.
comm_use.O-Z.xml.tar.gz	Dec. 13, 2017, 12:37 p.m.
non_comm_use.A-B.xml.tar.gz	Dec. 13, 2017, 12:18 p.m.
non_comm_use.C-H.xml.tar.gz	Dec. 13, 2017, 12:20 p.m.
non_comm_use.I-N.xml.tar.gz	Dec. 13, 2017, 12:28 p.m.
non_comm_use.O-Z.xml.tar.gz	Dec. 13, 2017, 12:25 p.m.

PubMed Central (PMC) query
query string

XML tags to take text from (leave empty for all)
for example: abstract, body

XML tags to ignore (leave empty for nothing)
for example: back, xref

Add class label to output file (optional)
for example: DOCSET_1

Take a random sample (optional)
for example: 200

☑ do not return empty XML extraction results (documents)

PMC article subset

○ free fulltext articles
● open access subset of free articles (OAC)

Guidelines

1. Do not specify nested XML tags to take text from or duplicates will occur!
2. Order of XML tags to take text from will determine the order of output text blocks.
3. Ignored XML blocks are deleted prior to the extraction of wanted blocks.
4. Attribute-value tag search is possible and has the following syntax, e.g.: `sec sec-type="results"`

Limitations

1. Due to NCBI API query limit we also enforce strict limiting: every user is limited to 1 query in 10 seconds, 5 queries per minute, 30 per hour, 200 per day.

© 2017 Vid Podpečan [download source from Github]
Jožef Stefan Institute & National Institute of Biology

Figure 1: A screenshot of the web application for search, extraction and download of free full-text articles from PMC.

To allow search and text mining on the complete free full-text document collection we have constructed a database which contains all available free full-text documents extracted from the compressed archives of XML files. On top of that, we provide a web application which offers PMC search, XML extraction and preprocessing, random sampling and export in the widely used *lndoc* format which can be read by OntoGen and CrossBee. Figure 1 shows a screenshot of the application in a browser window. The complete code of the web application is available under the MIT licence on GitHub: https://github.com/vpodpecan/pmcutils

Results

Queries to retrieve the articles from open access databases for several different domains that could be combined with plant immune signalling domain were prepared. An example of a query is shown in Figure 2.

- D1: (Arabidopsis OR Oryza OR Solanum OR Nicotiana OR maize OR cotton) AND ("defence" OR "defense" OR "immune" OR "hypersensitive" OR "hypersensitivity" OR "resistance" OR "susceptible" OR "susceptibility" OR "pathogen" OR "virus") NOT (cigarette OR smoking OR carcinoma OR cancer OR insulin OR diabetes OR "Parkinson's" OR mice OR rodents OR zebrafish OR teenagers)
- D2: (circadian) AND (rhythm OR clock)

Figure 2: An example of a query to retrieve the articles from open access databases.

Retrieved articles were processed Ontogen. On the basis of top keywords in OntoGen, the queries were refined. Several cycles of refining the queries were performed.

On the basis of SVM keywords obtained in OntoGen, the list of stop words was produced. Several cycles of refining the stop words were performed. An example of a stop words list is shown in Figure 3.

- fig
- figure
- table
- result
- mm
- al.
- cca
- cca.
- sequencing
- gene
- expression
- example
- use
- source
- method
- approach
- percentage
- al
- patients
- mice
- genes
- proteins
- mirnas
- rna
- transcriptional
- snp

- qtl
- genome
- cell
- root
- strains
- mutant

Figure 3: An example of a stop words list for OntoGen.