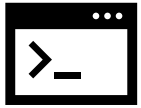
TEXAS
The University of Texas at Austin

Softcite:
Automatic Extraction of Software Mentions from Research Literature

Caifan Du[1], James Howison[1], Patrice Lopez[2]

[1] School of Information, University of Texas at Austin
[2] SCIENCE-MINER

SciNLP @AKBC2020



Software now becomes the first-class citizen in research_

But we cannot access them like retrieving research literature…

If we can extract mentions of software from research literature_

We will be able to:

→ Identify and retrieve software used in research

→ Understand and replicate research workflow

→ Give credit to researchers who write software

Softcite Dataset

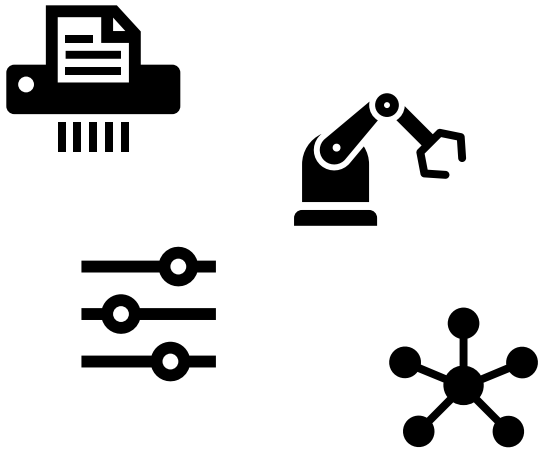
- 5,553 research articles in biomedicine & economics
- 5,210 software mentioned along with their attributes (including *publisher*, *version*, *URL*, and if the software was *used*)
- 8, 398 annotations in total

Softcite Dataset

```
<text xml:lang="en">
  <body>
    <p>All the analysis was performed in the <rs cert="1.0" resp="#annotator12"
    subtype="used" type="software" xml:id="f33d05cff5-software-0">MATLAB</rs> environment
    (<rs corresp="#f33d05cff5-software-0" resp="#curator" type="publisher">The MathWorks</
    rs>, Natick, MA) using <rs corresp="#f33d05cff5-software-1" resp="#curator"
    type="publisher">OMLAB</rs> software (<rs cert="1.0" resp="#curator" subtype="used"
    type="software" xml:id="f33d05cff5-software-1">OMtools</rs>, downloadable from <rs
    corresp="#f33d05cff5-software-1" resp="#curator" type="url">http://www.om lab.org</rs>).
    Eye position was sampled directly; it was prefiltered using a low-pass filter with a
```

Figure 1: Softcite TEI/XML corpus

Software Entity Recognition



Supervised Learning

- CRF
- BiLSTM-CRF with GloVes embeddings
- BiLSTM-CRF with GloVes and Elmo embeddings
- BERT-base-en+CRF
- SciBERT+CRF

Available at
https://github.com/ourresearch/software-mentions

Software Entity Disambiguation

Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W325–W329***
doi:10.1093/nar/gkm303

taveRNA: a web suite for RNA algorithms and applications

Cagri Aksay[1], Raheleh Salari[1], Emre Karakoc[1], Can Alkan[2] and S. Cenk Sahinalp[1,*]

[1]Lab for Computational Biology, SFU, Canada and [2]Department of Genome Sciences, University of Washington

Received January 31, 2007; Revised April 3, 2007; Accepted April 14, 2007

ABSTRACT

We present taveRNA, a web server package that hosts three RNA web services: alteRNA, inteRNA and pRuNA. alteRNA is a new alternative for RNA secondary structure prediction. It is based on a dynamic programming solution that minimizes the sum of energy density and free energy of an RNA structure. inteRNA is the first RNA–RNA interaction structure prediction web service. It also employs a dynamic programming algorithm to minimize the free energy of the resulting joint structure of the two interacting RNAs. Lastly, pRuNA is an efficient database pruning service; which given a query RNA, eliminates a significant portion of an ncRNA database and returns only a few ncRNAs as potential regulators. taveRNA is available at http://compbio.cs.sfu.ca/taverna.

INTRODUCTION

Until recently RNA was thought to have only two functions: (i) primarily as an information transmitter between DNA and proteins in the form of a messenger

Regulatory ncRNAs that are generally responsible for regulating gene expression exhibit an exact or partial complementarity to their target mRNAs. Their interaction forms a complex that consists of several non-contiguous helical segments which prevent ribosomal access to the target mRNA. Generally, regulatory ncRNAs contain one or more stem loop structures that are (almost) complementary to specific sequences in the target mRNAs. Interaction with a target RNA is either initiated at such a loop structure of the antisense RNA and a loop structure from the target (forming kissing loop pairs) or between a loop structure and a single-stranded segment of the complementary RNA.

As the number of ncRNAs and in particular regulatory RNAs increase it has become of crucial importance to establish software tools that can help identify their functionality. For this purpose we introduce taveRNA, a web-based computational tool set that can help identify structure and functionality of ncRNA molecules. taveRNA involves tools whose algorithmic foundations were developed by Simon Fraser University's Lab for Computational Biology over the past few years. The tools aim to solve the following key problems:

1. RNA secondary structure prediction problem, which

SciNLP @AKBC2020

TAVERNA

Type: **software**

Raw name: **taveRNA**

Creator: **the Lab for Computational Biology at Simon Fraser University**

conf: *0.748*

taveRNA is a **software suite** for **RNA/DNA** secondary structure. It is developed in the laboratories for **computational biology** of the School of **Computing Science** at the **Simon Fraser University**. The suite is composed by alteRNA, for RNA density fold computing, inteRNA, for RNA-RNA interaction prediction, **piRNA**, for predicting the joint partition function, equilibrium concentration, ensemble energy, and **melting temperature** for two RNA sequences, pRuNA, a sequence based **pruning** RNA interaction **search engine**, and smyRNA, a **platform independent C** program novel ab initio **ncRNA** finder.

Wikidata statements

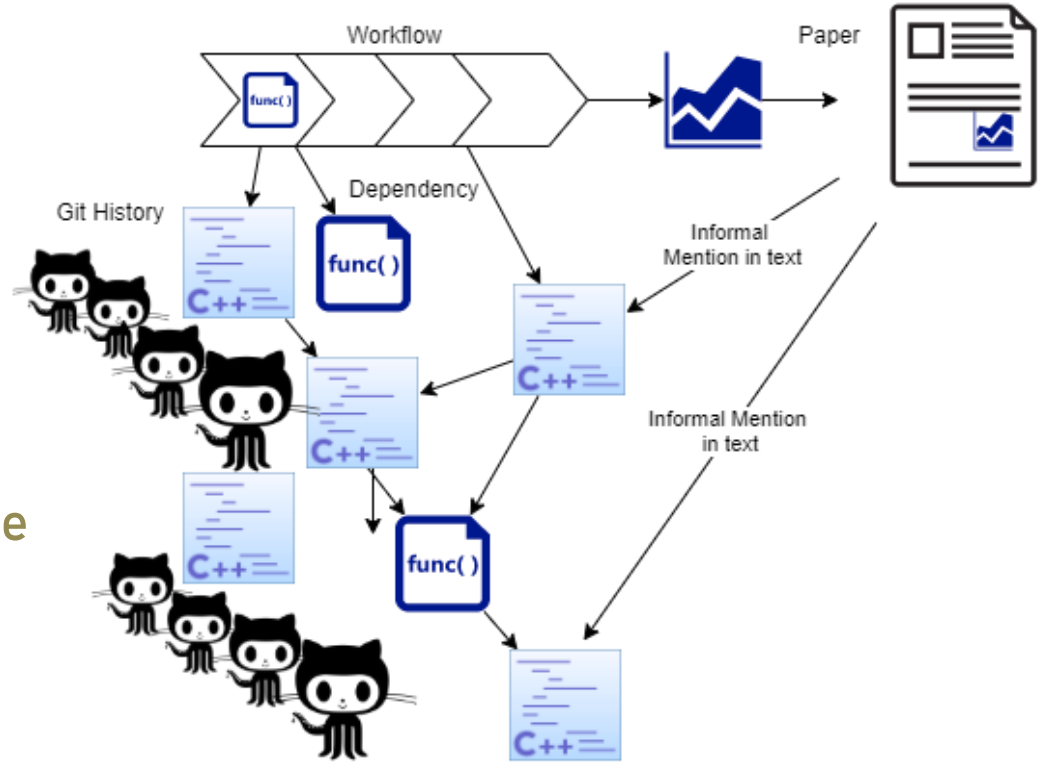
official website	http://compbio.cs.sfu.ca/taverna
use	Science
use	Bioinformatics
instance of	Software

References:

Software Knowledge Base

Application:

Link to software dependency data, software repository activity history, and research workflow objects, enabling analysis of domain software use and dependency risks



SciNLP @AKBC2020

THANKS!

This work is supported by
Alfred P. Sloan Foundation
Digital Science Program

Contact Info

C.Fan Du | cfdu@utexas.edu