A global bio-concept network for biomedical knowledge discovery tasks

Weize Xu

Huazhong Agriculture University

May 9, 2019

Outline

1. Data
2. Method
3. Result
4. Discussion
5. Reference

PubTator bioconception database

Download from NCBI PubTator[1] FTP server:

- version: 2019/01/05
- 28,581,464 article's abstract records
- 251,421,160 entities annotations.

[1] Chih-Hsuan Wei et al. "PubTator: a web-based text mining tool for assisting biocuration". In: *Nucleic acids research* 41.W1 (2013), W518–W522.

Technical route

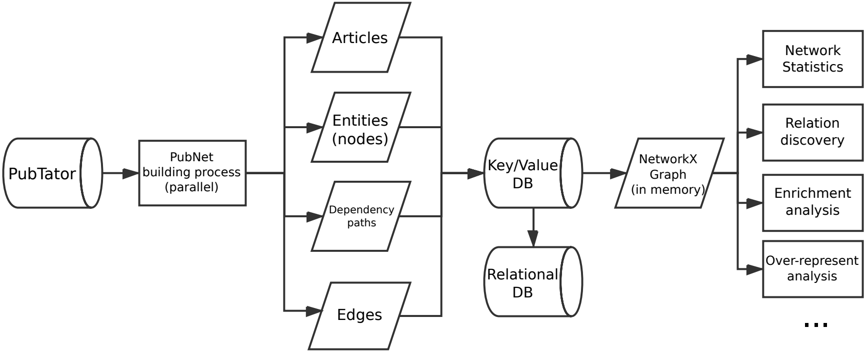

Figure: Workflow of bio-concept network building and data mining.

Network construct

- Dependency parse: Spacy[2]
- Edge strength between concept *i* and *j*:

$$s_{ij} = \sum_{p \in Paths_{ij}} \frac{1}{1 + length(p)} \tag{1}$$

- LSM Key/Value database: Sqlite4
- Graph manipulation: NetworkX[3]

[2] Matthew Honnibal et al. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* (2017).

[3] Aric Hagberg et al. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Network statistic

- Visualization of a sub-network
- Network statistic

Applications

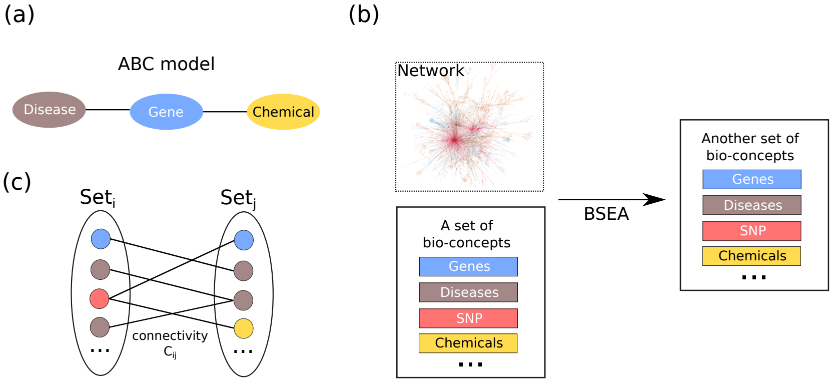

Figure: Schematic diagram about example applications of the network. (a) ABC model for drug discovery. (b) Bio-concept set enrichment analysis. (c) Connectivity over-represent analyze

BSEA results

Term ID	Term type	Describtion	Study ratio	Background ratio	p value
MESH:D011471	Disease	Prostatic Neoplasms	0.739130	0.033561	7.102240e-21
MESH:D009369	Disease	Neoplasms	0.869565	0.156557	8.499019e-14
MESH:D011470	Disease	Prostatic Hyperplasia	0.347826	0.007062	2.710133e-12
MESH:D007938	Disease	Leukemia	0.478261	0.031950	3.328331e-11
CHEBI:33704	Chemical	α-amino acid	0.695652	0.120383	2.050363e-10
MESH:D001943	Disease	Breast Neoplasms	0.521739	0.054373	5.103941e-10
MESH:D009223	Disease	Myotonic Dystrophy	0.304348	0.008640	7.715694e-10
MESH:D001523	Disease	Mental Disorders	0.434783	0.035047	2.089974e-09
MESH:D007713	Disease	Klinefelter Syndrome	0.173913	0.000729	2.374418e-09
MESH:D000596	Chemical	Amino Acids	0.434783	0.037761	4.264456e-09

Table: Top 5 BSEA results

Discussion

- Data source and validation
- Semantic information of connections
- Web services and architecture for more large scale graph analysis

Implementation

The codes in this paper used for construct the network and data mining is implemented as a Python package named PubNet. All data and codes in this paper will be open-source for academical usage.

Please feel free to contact me to do some awesome work together!

Reference

Hagberg, Aric, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Honnibal, Matthew and Ines Montani. “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. In: *To appear* (2017).

Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu. “PubTator: a web-based text mining tool for assisting biocuration”. In: *Nucleic acids research* 41.W1 (2013), W518–W522.