Drug knowledge discovery

by using BioNLP and tensor or matrix decomposition

Mina Gachloo

M_Gachloo@yahoo.com
Huazhong agriculture university

What is drug-related knowledge discovery?

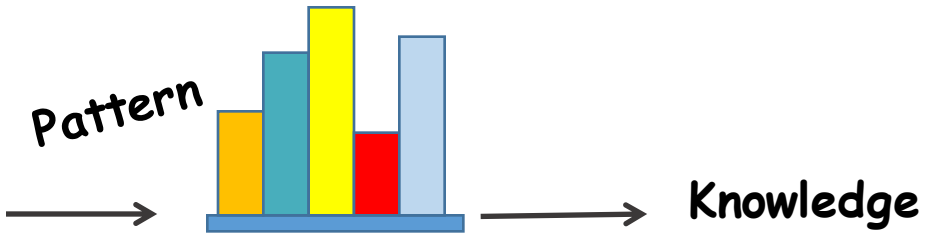
Drug-related knowledge discovery is process of discovering new knowledge which reflects novel drug-drug interaction, drug-disease or drug-indication linkage and drug-target identification.

With novel knowledge:

- Better understanding molecular basis of drug efficacy
- Focus on the application scenario of new drug discovery
- Drug repurposing or drug development.

Drug related-knowledge discovery

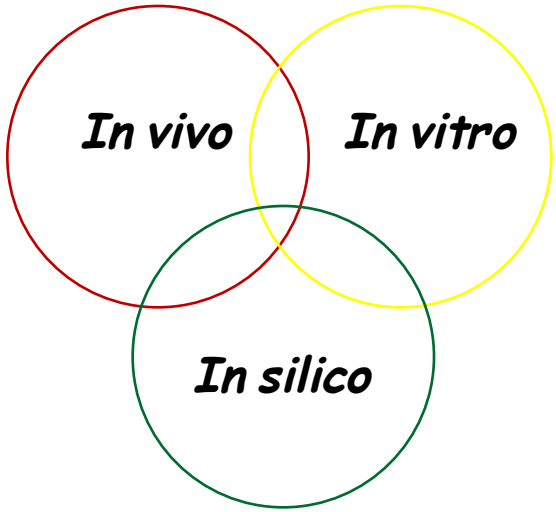
The main knowledge discovery pattern for drug related-knowledge discovery is predication of the relations between drug or other molecular and social entities. Computational approaches have combined the information from different source to provides relationships between drugs, *targets, disease and targeted genes.*



Drug related-knowledge discovery

Experimental methods:

- *In vivo*
- *In vitro*
- *In silico*



In silico methods for drug knowledge discovery

In silico method is a computational way to perform knowledge inference and mainly to perform knowledge discovery by data mining with less time-consumption.

In silico methods for drug knowledge discovery

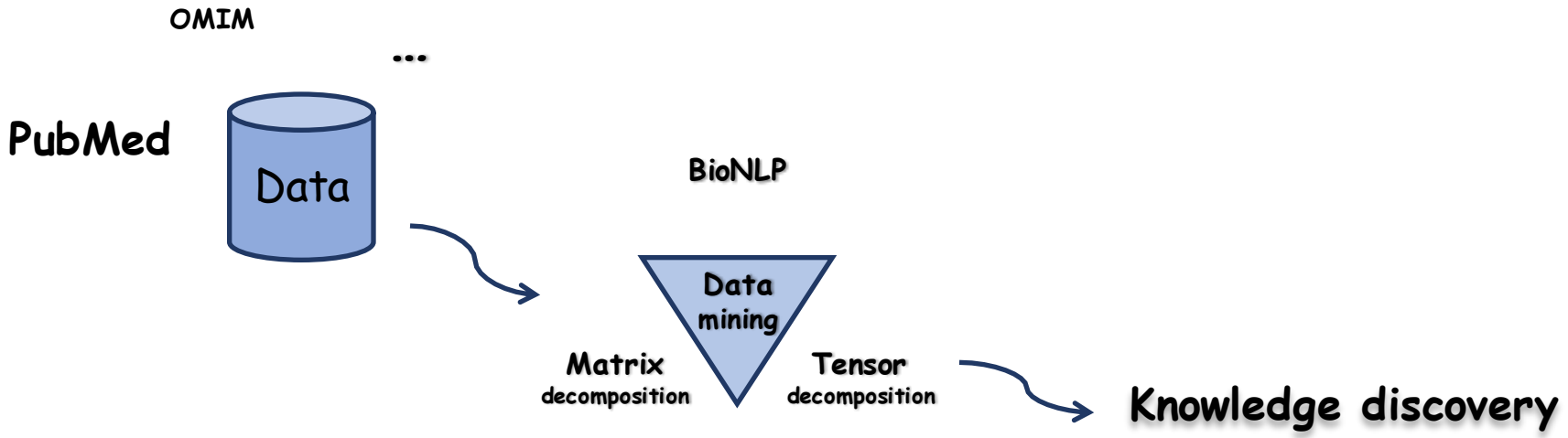
In silico methods including:

- ✓Machine learning
- ✓Molecular docking
- ✓ Pharmacophore
- ✓ Structure-activity relationship (SAR)
- ✓ Quantitative structure activity relationship (QSAR)
- ✓ And combination methods.

In silico methods for drug knowledge discovery

Two typical *in silico* methods:

- Biomedical Natural language processing (BioNLP)
- Matrix or Tensor decomposition



BioNLP

BioNLP is the application of natural language processing methods on biomedical *such as:*

1. *Relation extraction between protein-protein or drug- drug interactions.*
2. *Molecular entity recognition*

BioNLP

Three kinds of text resources *such as:*

1. *Large scale curation data (PubMed, OMIM ,…)*
2. *Small scale corpora*
3. *Heterogeneous data (Multi-omics)*

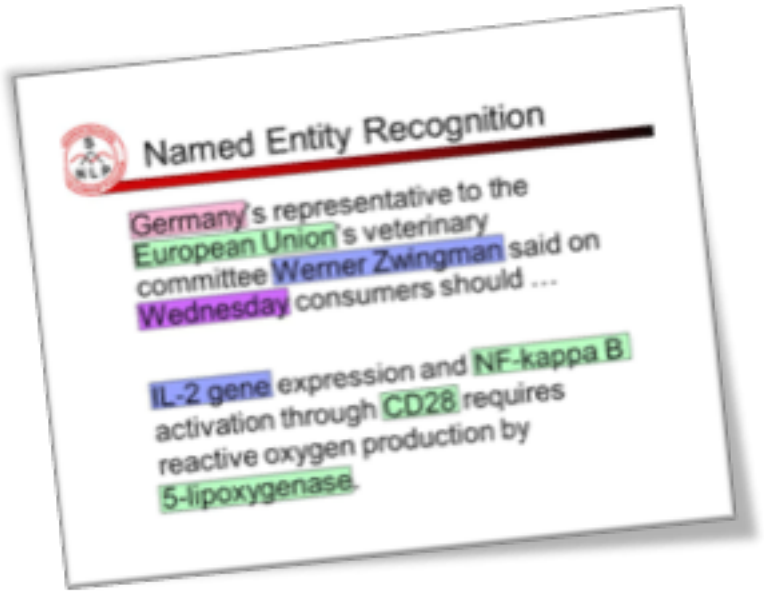
PubMed

Has long been the main text resources for BioNLP community to collect references and abstracts on life sciences and biomedical topics.

There were several text resources serving for drug-related knowledge discovery.

NER tools

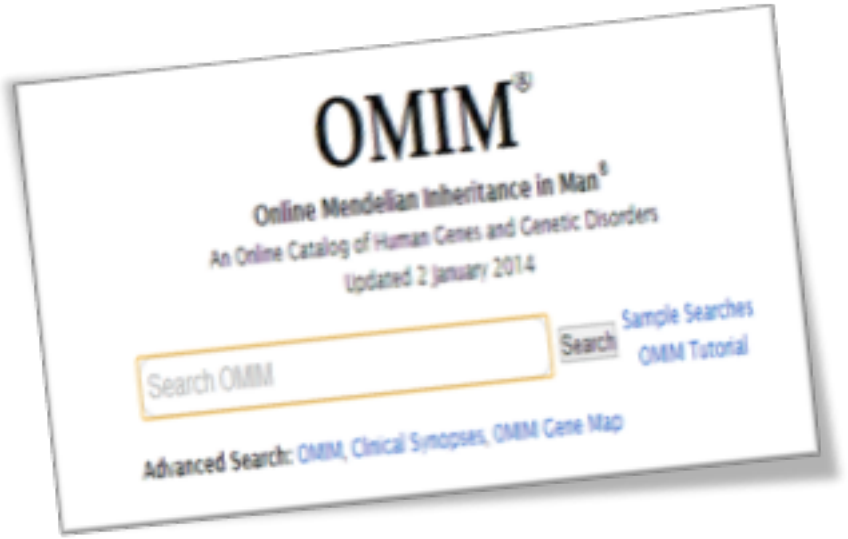
NER tools were developed among BioNLP community among dozens of popularized NER tools, tmChem, DNrom, GNormPlus, tmVar were regarded as successful representative of tools recognizing chemical, disease, gene and variation.



Online Mendelian Inheritance in Man(OMIM)

As a popular knowledge base of human genes and genetic disorders.

OMIM for drug mechanism.



Clinicaltrials.gov

Besides PubMed text and OMIM, Clinicaltrails.gov was a representative of electronic health record(EHR) text resource, which was established in 1990, contains different information about medical studies in human volunteers and the open access policy made it widely used.

Electronic Health Record(EHR)

EHR data is a popular sources information of clinical and transnational research for drug repurposing.

Annotation corpora

Annotation corpus is crucial to BioNLP, which cloud help to retrieve and extract information from biomedical text and also provide standard data for repeatable training and evaluation of BioNLP.

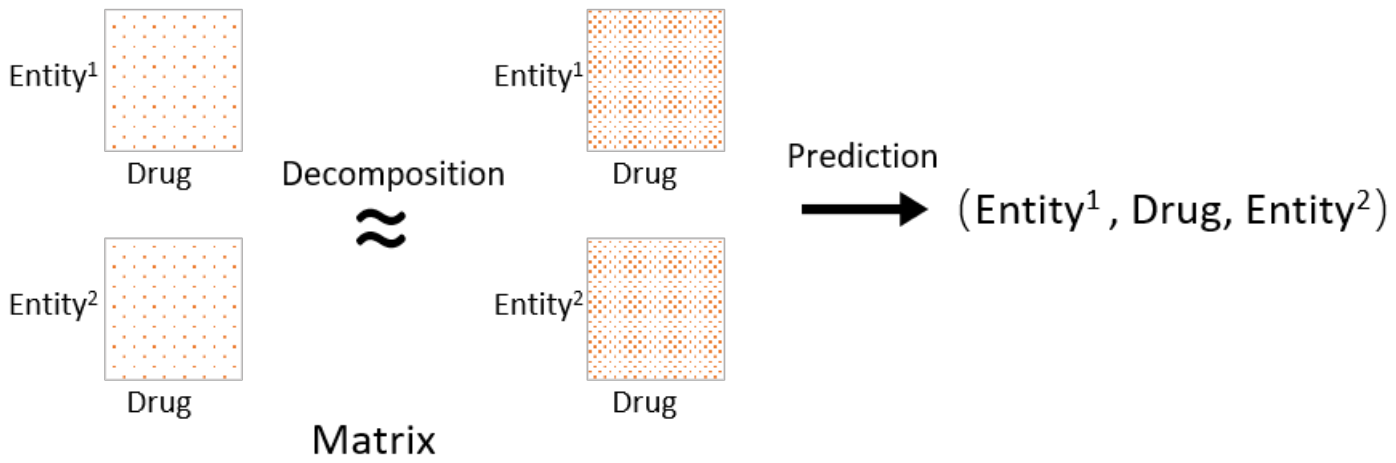
Multi-omics

Emergency of Multi-omics, the integration of text with genome, or protome data attracted attention from a cross disciplinary view for the purpose of drug-gene linking discovery.

Matrix decomposition method

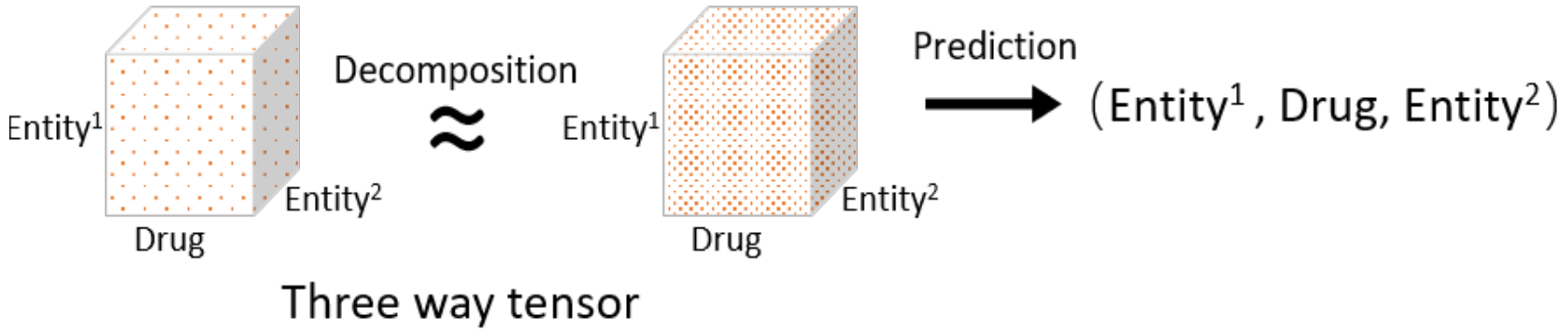
Matrix factorization or decomposition are important techniques for extraction information from a matrix or tensor.

The computational decomposition result of a matrix or tensor led to a-so-called low rank approximation of the original one and made a basis for novel link discovery.



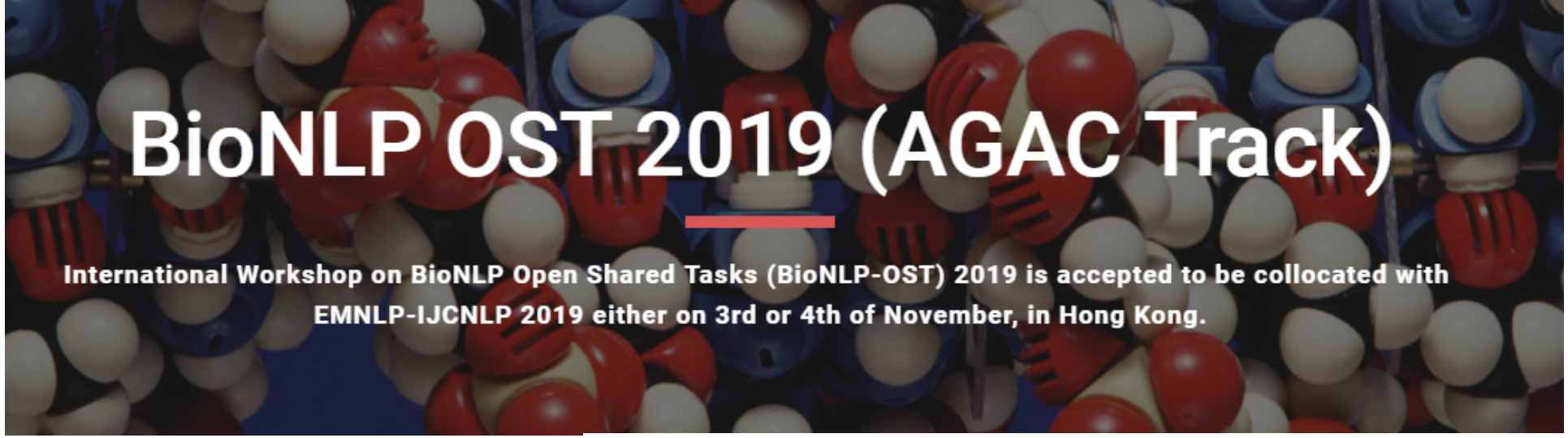
Tensor decomposition method

Tensor incorporated multidimensional array of numerical data and was applied in various machine learning tasks. Tensor decomposition extract a low rank approximation of drug data, but withholding more complex data structure.



AGAC

AGAC track provides an active gene annotated corpus (AGAC) and aims to extract mutation, disease knowledge from PubMed. The mutation-disease knowledge in this track is gene-mutation function change-disease, which not only contains the relationship between mutation and disease but also indicates the function change of the mutation, i.e., gain of function(GOF) and loss of function(LOF).



BioNLP OST 2019 (AGAC Track)

International Workshop on BioNLP Open Shared Tasks (BioNLP-OST) 2019 is accepted to be collocated with EMNLP-IJCNLP 2019 either on 3rd or 4th of November, in Hong Kong.

Timeline

11 Mar, 2019. Sample data (50 texts) release.

10 Apr, 2019. Training data (250 texts) release.

12 Jun, 2019. Testing data (1000 texts) release.

12-19, Jun, 2019. Evaluation period. (1 round results submission per day)

19 Jul, 2019. Workshop paper submission due

TBD. Notification of paper acceptance

TBD. Camera ready paper submission

3 or 4 Nov, 2019. Workshop to be collocated with EMNLP-IJCNLP 2019 (Hong Kong)

28 Mar, 2020. Special issue submission due

Data and Evaluation Codes

Data:

50 sample data. **Link.**

Full training data:

250 texts with NER and Rel annotation labels for Task 1 and Task 2. (**Click to download**)

250 "Gene;Function change;disease" links for Task 3. (**Click to download**)

Evaluation codes: Link

AGAC

AGAC track contain 3 different task:

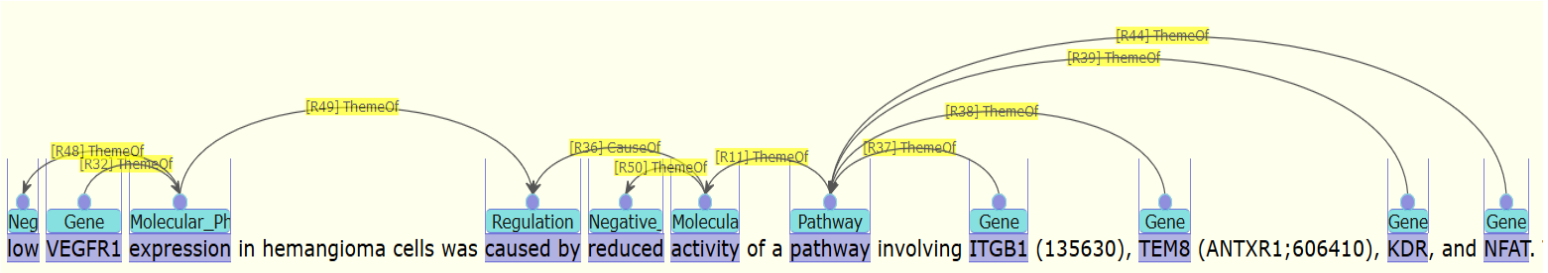
- ✓Trigger words NER
- ✓Themetic roles identification
- ✓Gene-function mutation-disease link discovery

Trigger words NER:

Recognize trigger words from PubMed abstract and annotated them with correspond AGAC labels or entities (Variation(Var), Molecular physiological activity(MPA), Interaction, Pathway, Cell physiological activity(CPA), Positive regulation(PosReg), Negative regulation(NegReg), Regulation, Disease, Gene, Protein, Enzyme).

Themetic roles identification:

It is to identify AGAC themetic roles(ThemeOf, CauseOf) between trigger words.



Gene-function mutation-disease link discovery

Extract the gene-(mutation)-function change-biology function or disease link.

Thank you