武漢大學
Wuhan University

Predicting drug-disease associations based on machine learning methods

Xiang Yue (岳翔)

Supervisor：Wen Zhang (章文)

Biomedical Big Data Mining Lab, Wuhan University

Lab site: http://bioinfotech.cn/

Homepage: https://xiangyue9607.github.io/

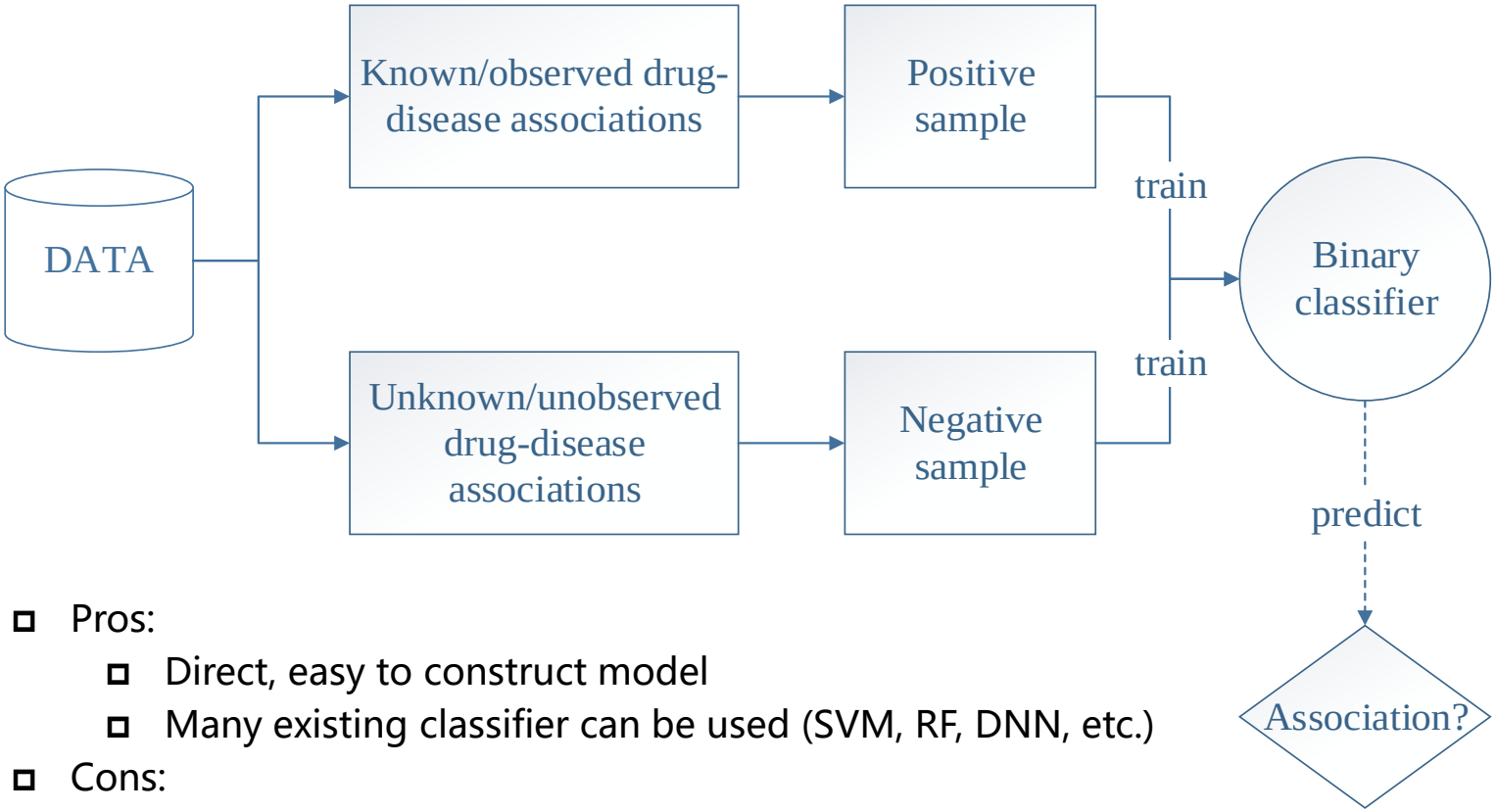
Background

- Drug-disease associations (What?):
  - drug indications (therapeutic functions)
  - other mechanisms (side effects, etc.)
- Mining drug-disease associations (Why?):
  - high incidence of disease VS time-consuming & expensive drug discovery
  - Identify potential drug therapeutic functions: Precision Medicine
  - Help drug repositioning
- Mining drug-disease associations (How?):
  - Traditional wet experiments
  - Computational methods:
    - Text Mining from literature
    - Statistics and machine learning-based prediction model

Background

- Drug-disease associations (What?):
  - drug indications (therapeutic functions)
  - other mechanisms (side effects, etc.)
- Mining drug-disease associations (Why?):
  - high incidence of disease VS time-consuming & expensive drug discovery
  - Identify potential drug therapeutic functions: Precision Medicine
  - Help drug repositioning
- Mining drug-disease associations (How?):
  - Traditional wet experiments
  - **Computational methods:**
    - Text Mining from literature
    - **Statistics and machine learning-based prediction model**

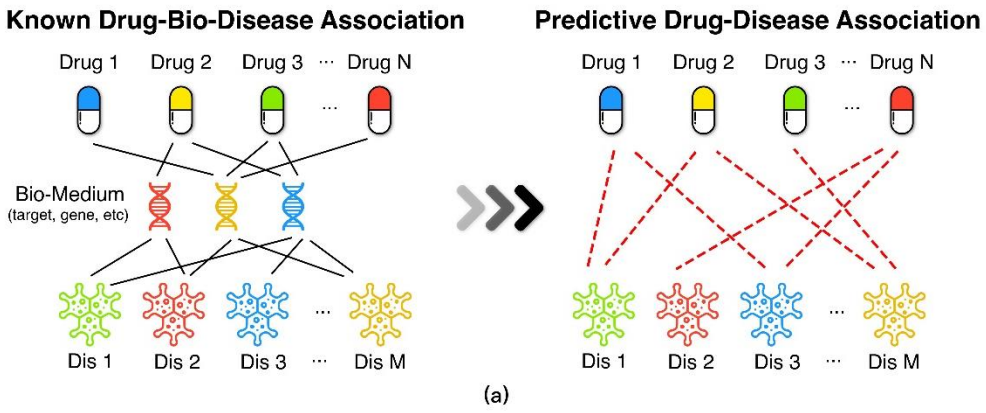
Binary Classification Task



- Pros:
  - Direct, easy to construct model
  - Many existing classifier can be used (SVM, RF, DNN, etc.)
- Cons:
  - How to construct training dataset (How to solve imbalance problem)?
  - Unobserved associations ≠ Negative samples

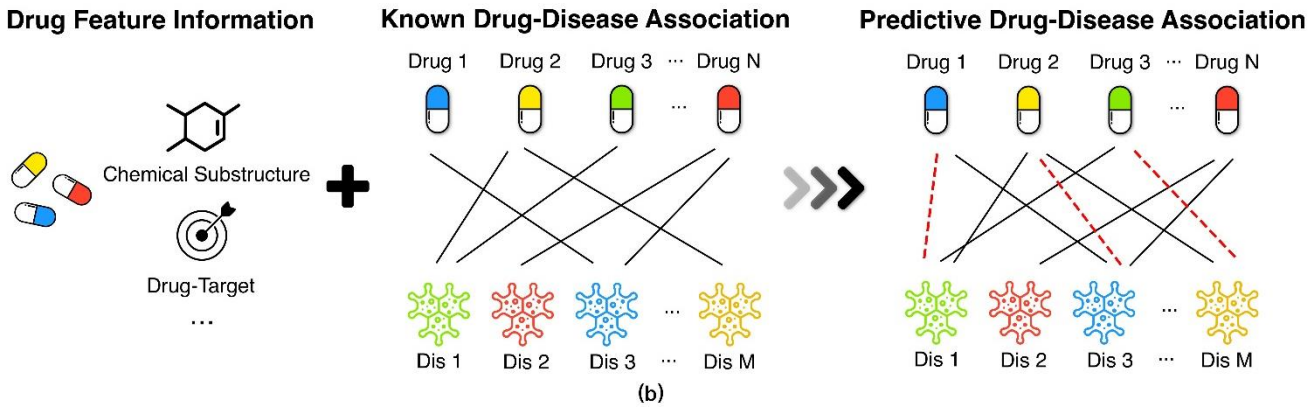
(Y. Wang *et al.*, 2013, M. Oh *et al.*, 2014, H. Moghadam *et al.*, 2016)

Network Inference Task



(a: L. Wang *et al.*, 2014, L. Yu *et al.*, 2015)

- Pros:
  - More interpretable
  - Avoid imbalance problem
- Cons:
  - How to describe the data point relations in the network?
  - How to handle heterogenous?
  - Data sparseness problem

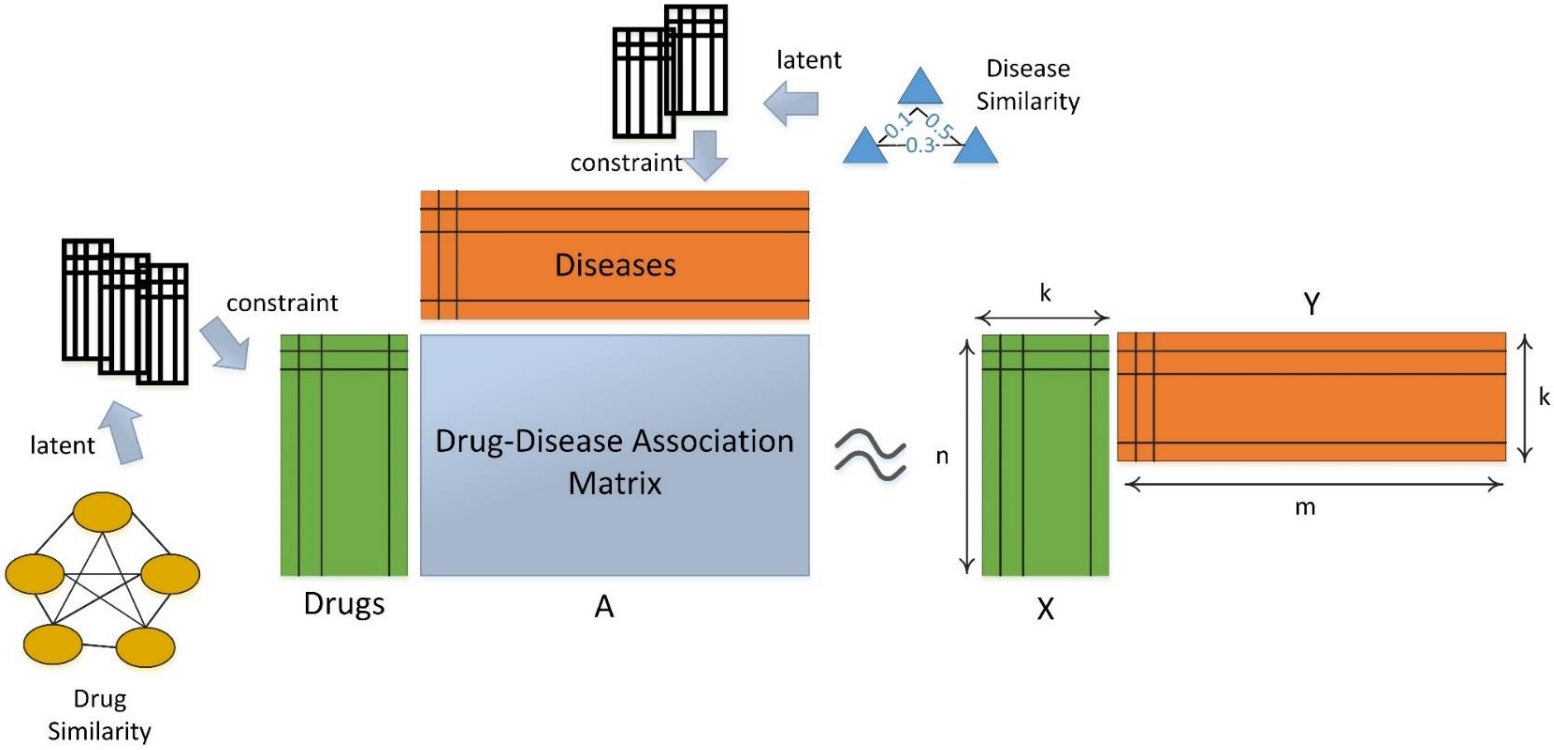


(b: Y.F. Huang *et al.*, 2013, W. Wang *et al.*, 2014, V. Martinez *et al.*, 2015, H. Wang *et al.*, 2015, X. Liang *et al.*, 2017)

Database

- Drug-disease associations:
  - Comparative Toxicogenomics Database (CTD): http://ctdbase.org/
  - ClinicalTrials.gov: https://www.clinicaltrials.gov/
- Drug features:
  - DRUGBANK (target, enzyme, drug-drug interaction, transporter, etc.): https://www.drugbank.ca/
  - PubChem (substructures): https://pubchem.ncbi.nlm.nih.gov/
  - SIDER (side effect): http://sideeffects.embl.de/
  - KEGG PATHWAY (pathway): https://www.genome.jp/kegg/pathway.html
- Disease features:
  - Medical Subject Headings (MeSH): https://meshb.nlm.nih.gov/
  - Online Mendelian Inheritance in Man (OMIM) (disease related genes): https://omim.org/

How to incorporate drug & dis. features?



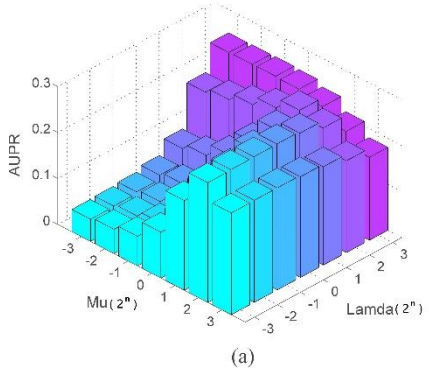
Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations by using similarity constrained matrix factorization, ***BMC Bioinformatics***, June 2018, DOI:10.1186/s12859-018-2220-4

Experiments

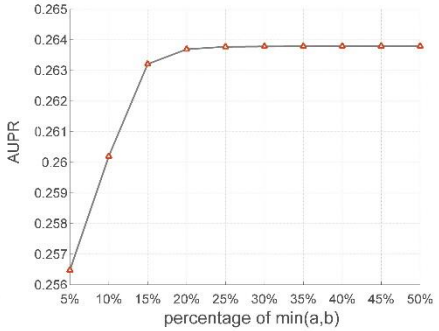
	AUPR	AUC	SN	SP	ACC	F
Substructure	0.2644	0.8737	0.3329	0.9795	0.9632	0.3130
Target	0.1947	0.8410	0.2751	0.9751	0.9575	0.2456
Pathway	0.2582	0.8706	0.3435	0.9771	0.9611	0.3079
Enzyme	0.2496	0.8671	0.3331	0.9768	0.9606	0.2990
Drug interaction	0.2638	0.8734	0.3505	0.9769	0.9611	0.3120

Methods	AUPR	AUC	SN	SP	ACC	F
PREDICT	0.1507	0.9020	0.3414	0.9929	0.9915	0.1437
SCMFDD-Che-GS	0.3141	0.9005	0.3663	0.9988	0.9974	0.3753
SCMFDD-Che-Phen	0.3153	0.9038	0.3678	0.9988	0.9974	0.3769
SCMFDD-SE-GS	0.3157	0.9082	0.3663	0.9988	0.9974	0.3753
SCMFDD-SE-Phen	0.3176	0.9109	0.3678	0.9988	0.9974	0.3769
SCMFDD-GP-GS	0.3210	0.9129	0.3720	0.9988	0.9975	0.3811
SCMFDD-GP-Phen	0.3224	0.9157	0.3714	0.9988	0.9975	0.3806
SCMFDD-GO-GS	0.3147	0.9035	0.3678	0.9988	0.9974	0.3769
SCMFDD-GO-Phen	0.3159	0.9065	0.3678	0.9988	0.9974	0.3769
SCMFDD-GW-GS	0.3249	0.9173	0.3389	0.9991	0.9977	0.3843
SCMFDD-GW-Phen	0.3284	0.9203	0.3776	0.9988	0.9975	0.3870

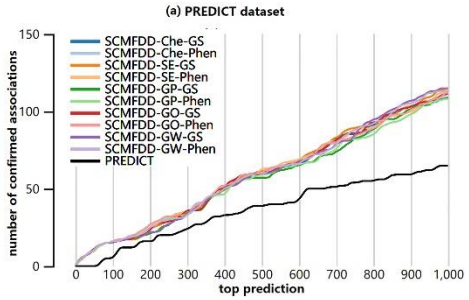
Methods	AUPR	AUC	SN	SP	ACC	F
TL-HGBI	0.0492	0.9584	0.1697	0.9999	0.9998	0.0840
SCMFDD	0.1500	0.9752	0.2136	0.9990	0.9990	0.0168

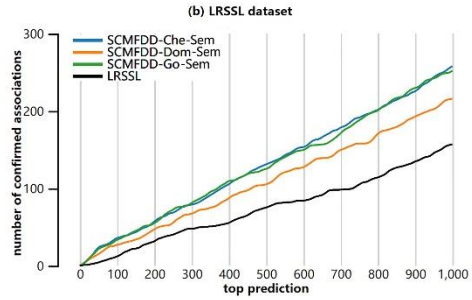


(a)



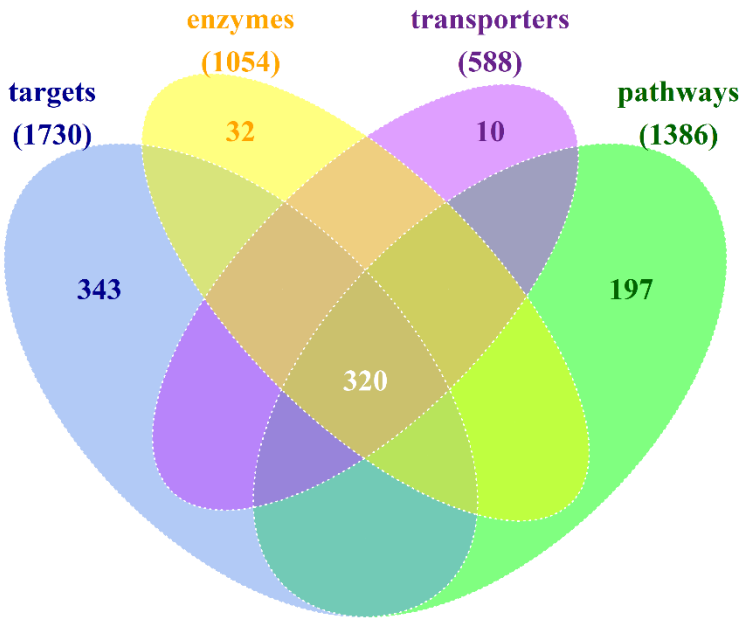
(b)





Methods	AUPR	AUC	SN	SP	ACC	F
LRSSL	0.1789	0.8250	0.2167	0.9989	0.9979	0.2018
SCMFDD-Che-Sem	0.2518	0.9020	0.2799	0.9993	0.9985	0.3030
SCMFDD-Dom-Sem	0.2673	0.9228	0.2851	0.9993	0.9985	0.3088
SCMFDD-Go-Sem	0.2585	0.9210	0.2897	0.9993	0.9985	0.3137

How to use less information to predict?

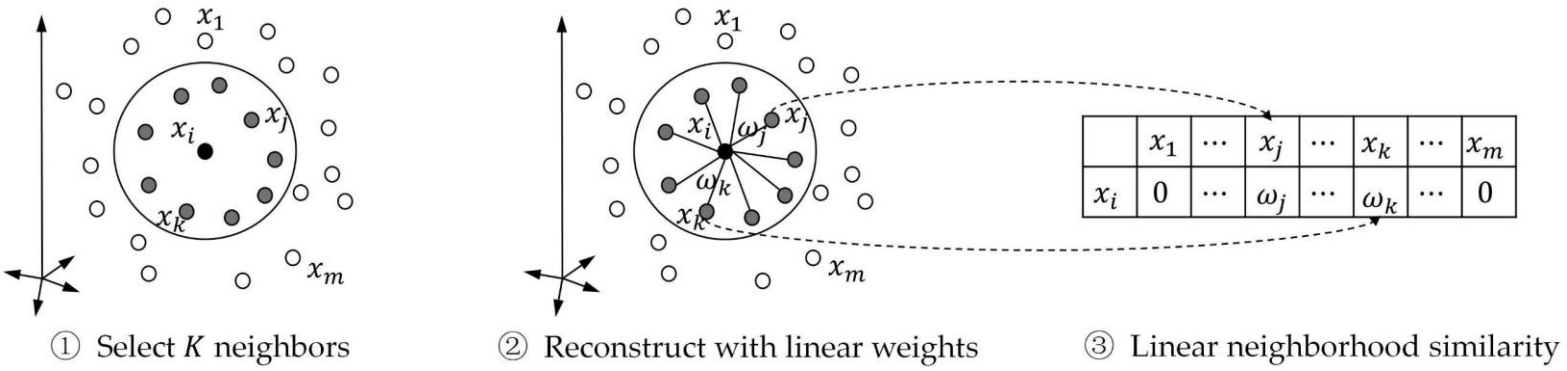


- Drug features can bring multiple information
- But not all features are available
- How to use less information to construct model?

Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations based on the known association bipartite network. ***2017 IEEE International Conference on Bioinformatics and Biomedicine(BIBM 2017)***, Kansan City, MO, USA, Nov 13 - Nov 16, 2017

Linear Neighborhood Similarity

- How to reconstruct every data point in the feature space?



- Objective function: minimize the reconstruct errors:

$$\min_{\omega_i} \varepsilon_i = \left\| x_i - \sum\nolimits_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} x_{i_j} \right\|^2 + \lambda \|\omega_i\|^2$$

$$s.t. \sum\nolimits_{i_j : x_{i_j} \in N(x_i)} \omega_{i,i_j} = 1,\ \omega_{i,i_j} \geq 0, j = 1, \cdots, K$$

Wen Zhang*, Xiang Yue, *et al*. A unified frame of predicting side effects of drugs by using linear neighborhood similarity. ***BMC systems biology***, 2017, 11(S6)

Wen Zhang*, Yanlin Chen, *et al*. Drug side effect prediction through linear neighborhoods and multiple data source integration. ***2016 IEEE International Conference on Bioinformatics and Biomedicine(BIBM 2016)***, ShenZhen, China, Dec 15-18, 2016

Fast Linear Neighborhood Similarity

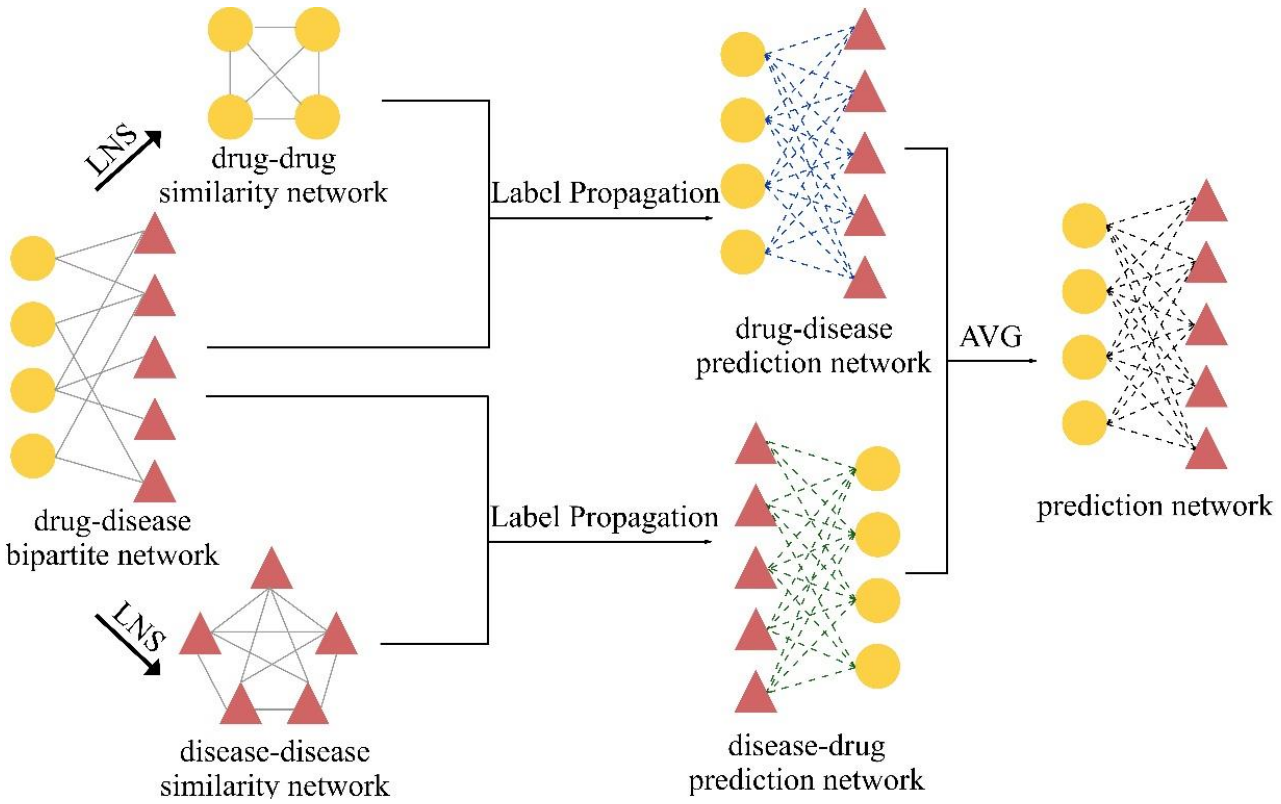
- Our proposed LNS has two problems:
  - Obtaining the weights of every data point needs to solve the optimization problem (e.g. 1000 data points need a few hours)
  - Many real-world problems have thousands and millions data points, existing framework could not apply on big data
- Consider all the data points at one time:

$$\min_{W} \|X - (C \odot W)X\|_F^2 + \lambda \|(C \odot W)e\|_1^2$$
$$\text{s.t.}\,(C \odot W)e = e, W \geq 0$$

$$C = (c_{ij}), if\ x_j \in N(x_i):\ c_{ij} = 1, else\ c_{ij} = 0, e = (1,1,1 \ldots, 1)^T$$

Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations based on the known association bipartite network. ***2017 IEEE International Conference on Bioinformatics and Biomedicine(BIBM 2017)***, Kansan City, MO, USA, Nov 13 - Nov 16, 2017

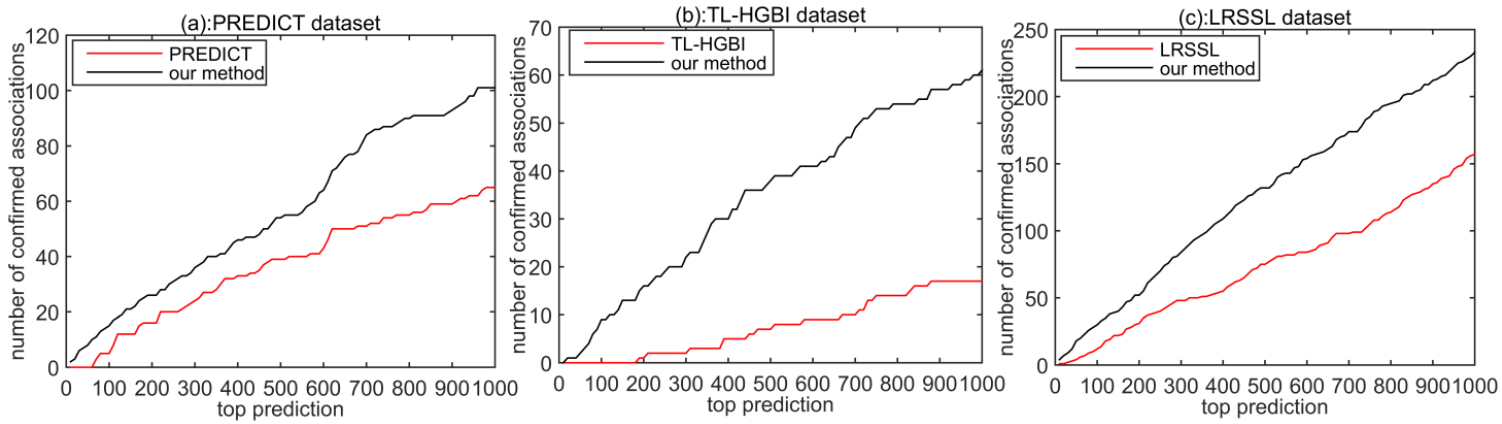
Network topological similarity inference method



Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations based on the known association bipartite network. ***2017 IEEE International Conference on Bioinformatics and Biomedicine(BIBM 2017)***, Kansan City, MO, USA, Nov 13 - Nov 16, 2017

Experiments

Methods	Datasets	AUPR	AUC	SEN	SPEC	PREC	ACC	F
PREDICT	PREDICT dataset	0.1507	0.9020	0.3414	0.9929	0.0914	0.9915	0.1437
Our method	PREDICT dataset	0.3376	0.9205	0.3678	0.9990	0.4624	0.9977	0.4022
TL-HGBI	TL-HGBI dataset	0.0492	0.9584	0.1697	0.9999	0.0571	0.9998	0.0840
Our method	TL-HGBI dataset	0.2631	0.9616	0.4032	0.9999	0.1658	0.9999	0.2349
LRSSL	LRSSL dataset	0.1789	0.8250	0.2167	0.9989	0.1988	0.9979	0.2018
Our method	LRSSL dataset	0.2693	0.9021	0.3078	0.9994	0.3757	0.9986	0.3384



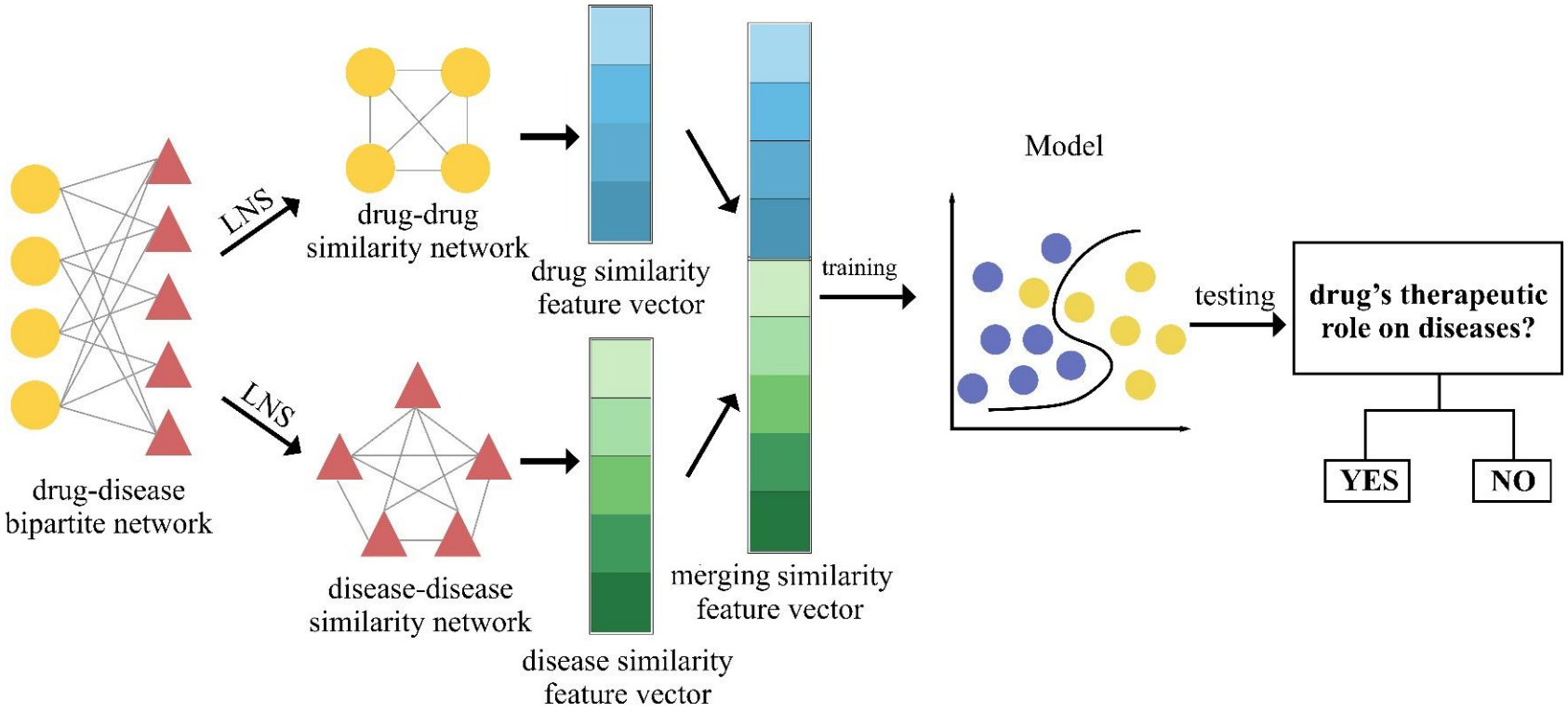
Methods	Datasets	AUPR	AUC	SEN	SPEC	PREC	ACC	F
Resource allocation	our dataset	0.1895	0.8408	0.2864	0.9738	0.2231	0.9564	0.2494
Our method	our dataset	0.2621	0.8709	0.3250	0.9805	0.3019	0.9640	0.3126
Resource allocation	PREDICT dataset	0.3212	0.8462	0.3580	0.9990	0.4362	0.9977	0.3923
Our method	PREDICT dataset	0.3376	0.9205	0.3678	0.9990	0.4624	0.9977	0.4022
Resource allocation	TL-HGBI dataset	0.0951	0.7747	0.1937	1.0000	0.1718	0.9999	0.1672
Our method	TL-HGBI dataset	0.2631	0.9616	0.4032	0.9999	0.1658	0.9999	0.2349
Resource allocation	LRSSL dataset	0.2094	0.8059	0.2734	0.9994	0.3483	0.9985	0.3025
Our method	LRSSL dataset	0.2693	0.9021	0.3078	0.9994	0.3757	0.9986	0.3384

How to further classify association type?

- Drug-disease associations:
  - drug indications (therapeutic functions)
  - other mechanisms (side effects, etc.)
- The proposed methods:
  - identify the potential drug-disease associations
  - fail to differentiate the association types

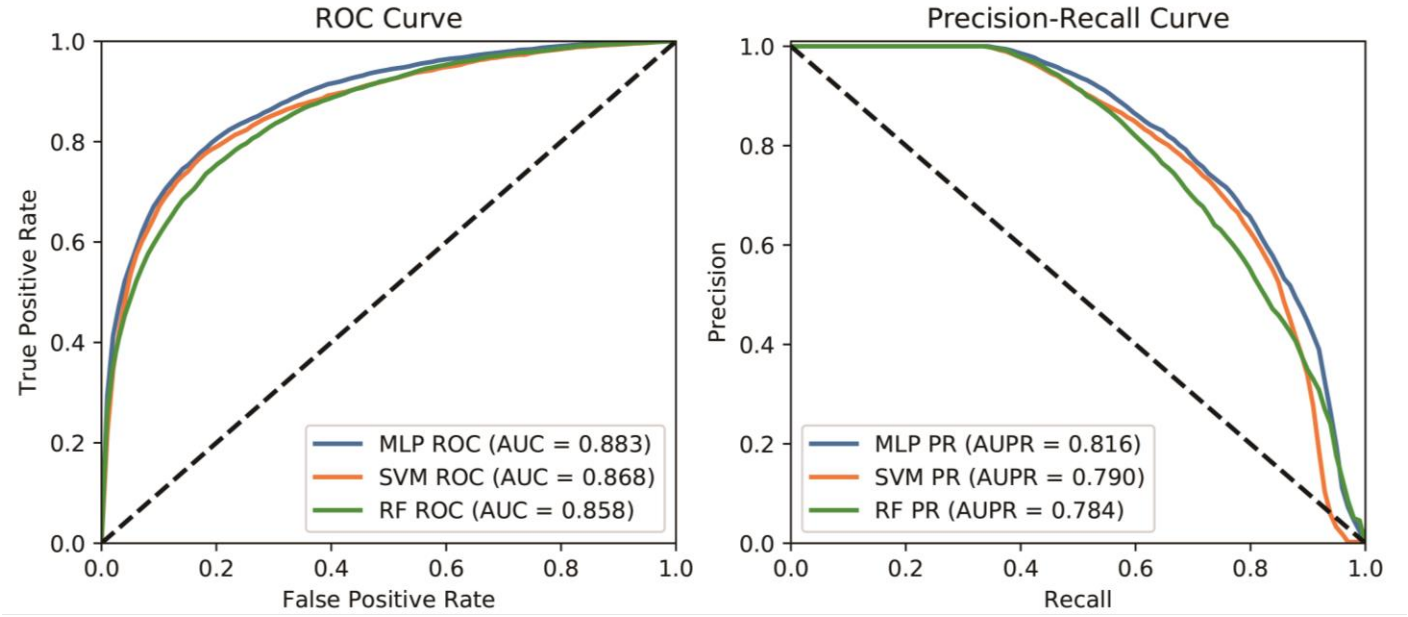
Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. ***Methods***, June 2018, DOI: 10.1016/j.ymeth.2018.06.001

How to further classify association type?



Wen Zhang*, Xiang Yue, *et al*. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. ***Methods***, June 2018, DOI: 10.1016/j.ymeth.2018.06.001

Experiment



Drug features	AUC	AUPR	PRE	REC	ACC	MCC	F
substructure	0.847	0.764	0.716	0.660	0.796	0.536	0.686
target	0.844	0.761	0.706	0.661	0.790	0.527	0.682
enzyme	0.845	0.760	0.693	0.674	0.788	0.524	0.682
pathway	0.850	0.768	0.719	0.652	0.795	0.534	0.683
drug-drug interaction	0.843	0.754	0.703	0.664	0.790	0.528	0.683
association profile	0.883	0.816	0.767	0.705	0.827	0.608	0.734

To get the predict result, please follow the instrument below:

1) Choose the drug or disease tab for the category of your input.

2) Choose the type of the input, MeSH ID, DrugBank ID, or PubChem CID, you can also input the name regardless of this option.

3) We suggest choosing a small value for the count of the result in order to get the result faster.

Database Don't know the MeSH ID?

MeSH DrugBank PubChem

Maximum Results

100

User Manual

Submit

Download Complete Data

Online Server:
http://bioinfotech.cn/SCMFDD/

Citation: Predicting Drug-Disease Associations by using the Similarity Constrained Matrix Factorization

Contact: Wen Zhang zhangwen@whu.edu.cn

Name: Clozapine | MeSH ID: D003024

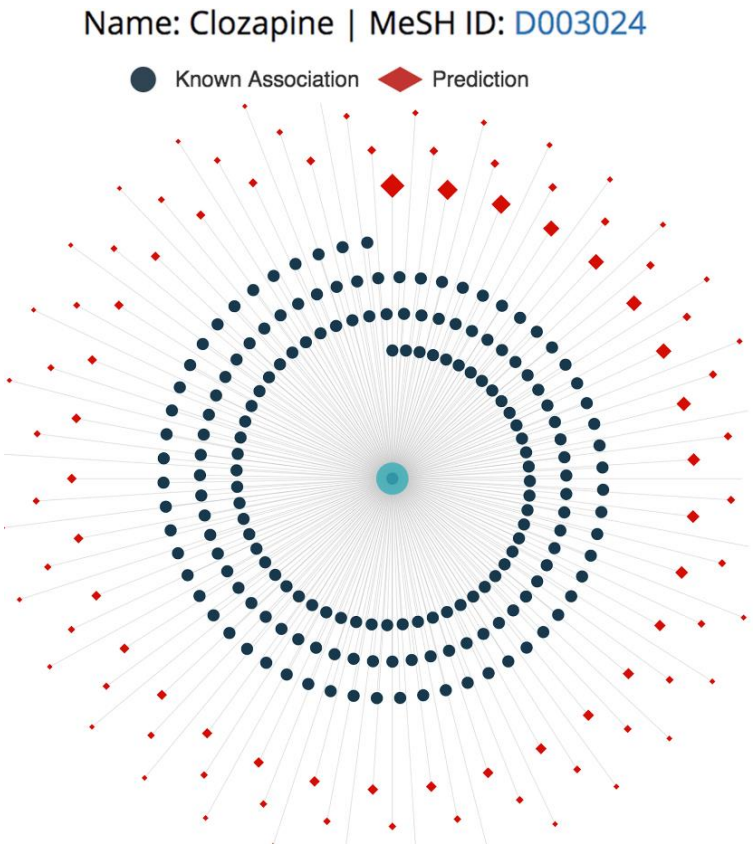
Prediction | Known Association | Visualization

Download CSV | Download PDF | Search:

Index	Disease Name	Disease MeSH ID	Score
1	Sleep Initiation and Maintenance Disorders	D007319	1
2	Anxiety Disorders	D001008	0.9117
3	Inappropriate ADH Syndrome	D007177	0.7434
4	Stress Disorders, Post-Traumatic	D013313	0.7267
5	Parkinson Disease, Secondary	D010302	0.7179
6	Memory Disorders	D008569	0.7123
7	Status Epilepticus	D013226	0.6312
8	Headache	D006261	0.6166
9	Torsades de Pointes	D016171	0.5953
10	Attention Deficit Disorder with Hyperactivity	D001289	0.5913

Show 10 entries

Previous 1 2 3 4 5 ... 10 Next



Results & Visualization

Online Server

Conclusion and future research

- Conclusion
  - Mining drug-disease associations is meaningful
  - Computational methods can accelerate the drug development and discovery
  - Known drug-disease associations are important for the prediction
- Future research
  - Incorporate more features into one framework (ensemble learning)
  - Develop more effective prediction models using less information
  - Pay more attention to classify the drug-disease association types (therapeutic or not)

Acknowledgement

Great appreciation and deep thanks for:

Prof. Wen Zhang


Yanlin Chen


Jingwen Shi


Canming Fang


Guifeng Tang


Xinrui Liu


Kanghong Jin


Jinghao Li


Feng Huang


Wenjian Wu


Weiran Lin


Yunqiu Zhang


Ding Zhang


Weitai Yang


Wenzheng Guo


Bolin Li


Xiaoting Lu


Siman Wang

Homepage: https://xiangyue9607.github.io/ Lab site: http://bioinfotech.cn/

Q&A