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

Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
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

Known/observed drug-disease associations → Positive sample → Binary classifier → predict

Unknown/unobserved drug-disease associations → Negative sample → Binary classifier

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)

Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
Network Inference Task

**Pros:**
- More interpretable
- Avoid imbalance problem

**Cons:**
- How to describe the data point relations in the network?
- How to handle heterogenous?
- Data sparseness problem

### Known Drug-Bio-Disease Association

- Drug 1, Drug 2, Drug 3, ..., Drug N
- Bio-Medium (target, gene, etc)
- Dis 1, Dis 2, Dis 3, ..., Dis M

### Predictive Drug-Disease Association

- Drug 1, Drug 2, Drug 3, ..., Drug N
- Dis 1, Dis 2, Dis 3, ..., Dis M

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

### Drug Feature Information

- Chemical Substructure
- Drug-Target ...

### Known Drug-Disease Association

- Drug 1, Drug 2, Drug 3, ..., Drug N
- Dis 1, Dis 2, Dis 3, ..., Dis M

### Predictive Drug-Disease Association

- Drug 1, Drug 2, Drug 3, ..., Drug N
- Dis 1, Dis 2, Dis 3, ..., Dis M

(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)

Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
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/
  ☐ SIDER (side effect): http://sideeffects.embl.de/

☐ 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?

# Experiments

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUPR</th>
<th>AUC</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL-HGBI</td>
<td>0.0492</td>
<td>0.9584</td>
<td>0.1697</td>
<td>0.9999</td>
<td>0.9998</td>
<td>0.0840</td>
</tr>
<tr>
<td>SCMFDD</td>
<td>0.1500</td>
<td>0.9752</td>
<td>0.2136</td>
<td>0.9990</td>
<td>0.9990</td>
<td>0.0168</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUPR</th>
<th>AUC</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRSSSL</td>
<td>0.1789</td>
<td>0.8250</td>
<td>0.2167</td>
<td>0.9989</td>
<td>0.9979</td>
<td>0.2018</td>
</tr>
<tr>
<td>SCMFDD-Che-Sem</td>
<td>0.2518</td>
<td>0.9020</td>
<td>0.2799</td>
<td>0.9993</td>
<td>0.9985</td>
<td>0.3030</td>
</tr>
<tr>
<td>SCMFDD-Dom-Sem</td>
<td>0.2673</td>
<td>0.9228</td>
<td>0.2851</td>
<td>0.9993</td>
<td>0.9985</td>
<td>0.3088</td>
</tr>
<tr>
<td>SCMFDD-Go-Sem</td>
<td>0.2585</td>
<td>0.9210</td>
<td>0.2897</td>
<td>0.9993</td>
<td>0.9985</td>
<td>0.3137</td>
</tr>
</tbody>
</table>
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?


Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
Linear Neighborhood Similarity

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

1. Select $K$ neighbors
2. Reconstruct with linear weights
3. Linear neighborhood similarity

- Objective function: minimize the reconstruct errors:

$$
\min_{\omega_i} \varepsilon_i = \left\| x_i - \sum_{i_j: x_j \in N(x_i)} \omega_{i,j} x_j \right\|^2 + \lambda \|\omega_i\|^2 \\
\text{s.t. } \sum_{i_j: x_j \in N(x_i)} \omega_{i,j} = 1, \ \omega_{i,j} \geq 0, j = 1, \ldots, K
$$


Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
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:

\[
\min_W \|X - (C \odot W)X\|_F^2 + \lambda \|(C \odot W)e\|_1^2 \\
s.t. (C \odot W)e = e, W \geq 0
\]

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


Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
Network topological similarity inference method


Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
## Experiments

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

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

How to further classify association type?


Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
**Experiment**

<table>
<thead>
<tr>
<th>Drug features</th>
<th>AUC</th>
<th>AUPR</th>
<th>PRE</th>
<th>REC</th>
<th>ACC</th>
<th>MCC</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>substructure</td>
<td>0.847</td>
<td>0.764</td>
<td>0.716</td>
<td>0.660</td>
<td>0.796</td>
<td>0.536</td>
<td>0.686</td>
</tr>
<tr>
<td>target</td>
<td>0.844</td>
<td>0.761</td>
<td>0.706</td>
<td>0.661</td>
<td>0.790</td>
<td>0.527</td>
<td>0.682</td>
</tr>
<tr>
<td>enzyme</td>
<td>0.845</td>
<td>0.760</td>
<td>0.693</td>
<td>0.674</td>
<td>0.788</td>
<td>0.524</td>
<td>0.682</td>
</tr>
<tr>
<td>pathway</td>
<td>0.850</td>
<td>0.768</td>
<td>0.719</td>
<td>0.652</td>
<td>0.795</td>
<td>0.534</td>
<td>0.683</td>
</tr>
<tr>
<td>drug-drug interaction</td>
<td>0.843</td>
<td>0.754</td>
<td>0.703</td>
<td>0.664</td>
<td>0.790</td>
<td>0.528</td>
<td>0.683</td>
</tr>
<tr>
<td>association profile</td>
<td>0.883</td>
<td>0.816</td>
<td>0.767</td>
<td>0.705</td>
<td>0.827</td>
<td>0.608</td>
<td>0.734</td>
</tr>
</tbody>
</table>
To get the predict result, please follow the instructions 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.

Search for Drug ID, Name... (EX: D003024 or clozapine)

Database
- MeSH
- DrugBank
- PubChem

Don't know the MeSH ID?

Maximum Results

Submit

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

Copyright 2017 BBDM Lab., Web Server developed by Wenjian Wu (Developer), Ruoqi Liu (Designer) and Xiang Yue (Chief).
Citation: Predicting Drug-Disease Associations by using the Similarity Constrained Matrix Factorization
Contact: Wen Zhang, zhengwen@whu.edu.cn
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

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

Presenter: Xiang Yue, Supervisor: Wen Zhang, BBDM-Lab, Wuhan Univ.
Q&A