Predicting Drug-Target Interaction for New Drugs Using Enhanced Similarity Measures and Super-Target Clustering

Jian-Yu Shi (NWPU)
Siu-Ming Yiu (HKU)
Yiming Li (HKU)
Henry C.M. Leung (HKU)
Francis Y.L. Chin (HKU)
神農

(shen nong; “divine farmer”)
Herbs

Shen Nong’s Organs

Interaction

Activity related to

Treatment

Disease
Drug

Protein

Interaction

Activity related to

Treatment

Schizophrenia (a disease)
Drug discovery: Predicting drug-target interaction is the key!
The prediction problem

Four scenarios:
1. Known drug, known target
2. New drug, known target
3. Known drug, new target
4. New drug, new target

**Input**

- **Drug similarity**
  - Drugs

- **Target similarity**
  - Targets

**Drug-target interaction**

- Drugs: $d_1, d_2, d_3, d_4$
- Targets: $t_1, t_2, t_3, t_4, t_5$

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$t_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$d_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$d_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Train a model for prediction

Problem with training data: missing interactions
Existing method #1: WNN-GIP

Weighted nearest neighbor – Gaussian interaction profile

(PloS One 2013)

Drug-target interaction

Biased!

Only uses positive samples to build the model
Existing method #2: KBMF2K

Kernelized Bayesian matrix factorization
(Bioinformatics 2012)

Drug-target interaction score matrix

Problem with training data: missing interactions

Hard to explain!

Drug similarity
Drug “latent feature”
Target “latent feature”
Target similarity

Drug similarity
Drug “latent feature”
Target “latent feature”
Target similarity
Limitations of the existing methods

WNNGIP and KBMF2K

- Missing interactions
- The similarity measure
  - Only based on the chemical structure of drugs and protein sequences of targets
Drug-target interaction prediction as probabilistic events
The neighbor idea

• A drug’s neighbors: the drugs most similar to it

• Predict a new drug’s behavior by its neighbors’ behavior
The probability

- Event A: to be predicted
  (New) drug $d$ interacts with target $t$
- Event B: the observation
  # of $d$'s neighbors interacting with target $t$

We calculate $Pr(A|B)$ by

$$Pr(A|B) = \frac{Pr(AB)}{Pr(AB) + Pr(A^C B)}$$

Probability of how likely $d$ interacts with $t$
given the observed number of interactions
of $d$'s neighbors with $t$
Our contribution #1

“Super-targets”
Cluster targets using similarities; Cluster = Super-target

\[ \begin{array}{ccccc}
  & t_1 & t_2 & t_3 & t_4 & t_5 \\
 d_1 & 0 & 1 & 0 & 0 & 0 \\
 d_2 & 1 & 0 & 1 & 1 & 1 \\
 d_3 & 0 & 0 & 1 & 0 & 1 \\
 d_4 & 0 & 0 & 0 & 0 & 1 \\
\end{array} \]

\[ \begin{array}{cc}
  & d_1 & d_2 & d_3 & d_4 \\
 t_1 & 1 & 0 \\
 t_2 & 1 & 1 \\
 t_3 & 0 & 1 \\
 t_4 & 0 & 1 \\
\end{array} \]

\( st = \text{super-targets} \)
Aim: \( S_X: \text{Prob}(d_2 \text{ interacts with } st_1) \)

\[ st = \text{super-targets} \]

\[
\begin{array}{c|c|c}
\text{d}_1 & \text{d}_2 & \text{d}_3 & \text{d}_4 \\
\hline
1 & 1 & 0 & 1
\end{array}
\]
Aim:

If we only use $S_x$, we are assuming all the targets in $st_1$ are equivalent.

$S_x$: $\text{Prob}(d_2 \text{ interacts with } st_1)$
For new drugs it is the same!

\[ S_x : \text{Prob}(d_2 \text{ interacts with } st_1) \]

\[ S_{Y/X} : \text{Prob}(d_2 \text{ interacts with } t_2 \mid d_2 \text{ interacts with } st_1) \]

\[ S_x S_{Y/X} : \text{Prob}(d_2 \text{ interacts with } t_2 \text{ in } st_1) \]

\[ st = \text{super-targets} \]

\[
\begin{array}{c|c|c}
   & st_1 & st_2 \\
\hline
   d_1 & 1 & 0 \\
   d_2 & 1 & 1 \\
   d_3 & 0 & 1 \\
   d_4 & 0 & 1 \\
\end{array}
\]

Aim:
Cluster targets using similarities; Cluster = Super-target

\[
\begin{array}{c|ccccc}
 & t_1 & t_2 & t_3 & t_4 & t_5 \\
\hline
d_1 & 0 & 1 & 0 & 0 & 0 \\
d_2 & 1 & 0 & 1 & 1 & 0 \\
d_3 & 0 & 0 & 1 & 0 & 1 \\
d_4 & 0 & 0 & 0 & 0 & 1 \\
\end{array}
\]

\[st = \text{super-targets}\]

\[
\begin{array}{c}
st_1 \\
st_2 \\
t_6
\end{array}
\]
Cluster targets using similarities; Cluster = Super-target

\[ st = \text{super-targets} \]

<table>
<thead>
<tr>
<th></th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>( t_4 )</th>
<th>( t_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( d_3 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( d_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

A new target could be clustered into one of the super-targets

For new drugs it is the same!
Our contribution #2

Enhanced similarity measures for drugs and targets
Existing similarity measures

Drugs: aligning the 2D chemical structures

Targets: aligning the protein sequences
They have low structural similarity (0.275) but share many targets.

* 2D chemical structures extracted from KEGG.
They have low structural similarity (0.275) but share many targets.

Non-structural similarity measures are needed!

* 2D chemical structures extracted from KEGG.
Anatomical Therapeutic Chemical Classification System

C03CA01

Hierarchical

Level 5: chemical substance
Level 4: therapeutic subgroup
Level 3: therapeutic subgroup
Level 2: therapeutic main group
Level 1: organ or system it acts on

Furosemide
Anatomical Therapeutic Chemical Classification System

D 05 B B 01  
D 05 A X 05

First two levels are the same!  
ATC code similarity = 2/5 = 0.4 > 0.275
Functional categories of proteins

- Non-structural
- Describing their **functions**
Our new similarity measure

Drugs

\[
\frac{\text{2D chemical structure similarity} + \text{ATC code similarity}}{2}
\]

Targets

\[
\frac{\text{protein sequence similarity} + \text{functional category code similarity}}{2}
\]
Using new similarity measures and “super-targets”

Our performance
AUC                                 AUPR

FP/(FP+TN)

A is better than B

Actual +ve  Actual -ve

Predicted +ve  TP  FP
       Predicted -ve  FP  TN

Predicted +ve  TP  FP
       Predicted -ve  FN  TN
AUC \quad \text{A is better than B} \quad \text{AUPR}

\begin{align*}
\text{FP}/(\text{FP}+\text{TN}) & \quad \text{TP}/(\text{TP}+\text{FN}) \\
\text{Actual +ve} & \quad \text{Actual -ve} \\
\text{Predicted +ve} & \quad \text{TP} \\
\text{Predicted -ve} & \quad \text{FP} \\
\end{align*}

Missing interactions:
Much more negative samples

When # of FP’s is big:
AUC is overly optimistic
## Overall performance

<table>
<thead>
<tr>
<th></th>
<th>Enzyme</th>
<th></th>
<th>Ion channel</th>
<th></th>
<th>GPCR</th>
<th></th>
<th>Nuclear receptor</th>
<th></th>
<th>Total running time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td></td>
</tr>
<tr>
<td>KBMF2K</td>
<td>0.812</td>
<td>0.287</td>
<td>0.802</td>
<td>0.245</td>
<td>0.840</td>
<td>0.347</td>
<td>0.810</td>
<td>0.354</td>
<td>115.4 min</td>
</tr>
<tr>
<td>WNN-GIP</td>
<td>0.861</td>
<td>0.280</td>
<td>0.775</td>
<td>0.233</td>
<td>0.872</td>
<td>0.311</td>
<td>0.839</td>
<td>0.456</td>
<td>190.9 min</td>
</tr>
<tr>
<td>Ours</td>
<td>0.812</td>
<td>0.385</td>
<td>0.811</td>
<td>0.367</td>
<td>0.875</td>
<td>0.414</td>
<td>0.871</td>
<td>0.533</td>
<td>5.5 min</td>
</tr>
</tbody>
</table>
With and without new similarity measures

<table>
<thead>
<tr>
<th></th>
<th>Enzyme</th>
<th></th>
<th>Ion channel</th>
<th></th>
<th>GPCR</th>
<th></th>
<th>Nuclear receptor</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
</tr>
<tr>
<td>Without new</td>
<td>0.805</td>
<td>0.332</td>
<td>0.776</td>
<td>0.296</td>
<td>0.854</td>
<td>0.304</td>
<td>0.860</td>
<td>0.476</td>
</tr>
<tr>
<td>With new</td>
<td>0.812</td>
<td>0.385</td>
<td>0.811</td>
<td>0.367</td>
<td>0.875</td>
<td>0.414</td>
<td>0.871</td>
<td>0.533</td>
</tr>
</tbody>
</table>
New drug, new target

- Remove known interactions from the data set to create “new” drugs and targets
- Consider if the removed interactions could be predicted
- The mis-prediction error measures the fraction of “new” drugs with a wrong prediction
New drug, new target

<table>
<thead>
<tr>
<th></th>
<th>Enzyme</th>
<th>Ion Channel</th>
<th>GPCR</th>
<th>Nuclear Receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBMF2K</td>
<td>0.774</td>
<td>0.600</td>
<td>0.654</td>
<td>0.600</td>
</tr>
<tr>
<td>WNN-GIP</td>
<td>0.931</td>
<td>0.600</td>
<td>0.692</td>
<td>0.600</td>
</tr>
<tr>
<td>Ours</td>
<td>0.657</td>
<td>0.500</td>
<td>0.500</td>
<td>0.600</td>
</tr>
</tbody>
</table>

The numbers are mis-prediction errors. The smaller the mis-prediction error, the better the performance.
Conclusions

• Non-structural-based similarities
• “Super-targets”

My e-mail: liym1018@hku.hk

Thank you for listening.
References

Image sources

- Schizophrenia patient artwork. [http://upload.wikimedia.org/wikipedia/commons/b/b2/Cloth_embroidered_by_a_schizophrenia_sufferer.jpg](http://upload.wikimedia.org/wikipedia/commons/b/b2/Cloth_embroidered_by_a_schizophrenia_sufferer.jpg)
- D2R. [http://en.wikipedia.org/wiki/Dopamine_receptor_D2#mediaviewer/File:Protein_DRD2_PDB_1l15.png](http://en.wikipedia.org/wiki/Dopamine_receptor_D2#mediaviewer/File:Protein_DRD2_PDB_1l15.png)
- Protein structure #1. [http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg](http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg)
- Protein structure #2. [http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg](http://upload.wikimedia.org/wikipedia/commons/8/86/Argonne's_Midwest_Center_for_Structural_Genomics_deposits_1,000th_protein_structure.jpg)

* All the images used are labeled for non-commercial reuse by Google image search.
Supplementary #1

Estimating $\Pr(A)$ and $\Pr(A^C)$

\[
\Pr [a(x, j) = 1] \approx \left[ 1 + \sum_{i=1}^{m} A(i, j) \right] / (m + 2);
\]
\[
\Pr [a(x, j) = 0] = 1 - \Pr [a(x, j) = 1]
\]

- Event A: (New) drug $d$ interacts with target $t$
- Event B: $c$ drugs in the set of $d$’s $K$ nearest neighbors interacts with target $t$
Event A: (New) drug $d$ interacts with target $t$

Event B: $c$ drugs in the set of $d$’s $K$ nearest neighbors interacts with target $t$
## Supplementary #2

All the methods with new similarity measures

<table>
<thead>
<tr>
<th></th>
<th>Enzyme</th>
<th></th>
<th>Ion Channel</th>
<th></th>
<th>GPCR</th>
<th></th>
<th>Nuclear Receptor</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
</tr>
<tr>
<td>KBMF2K</td>
<td>0.870</td>
<td>0.391</td>
<td>0.833</td>
<td>0.330</td>
<td>0.878</td>
<td>0.414</td>
<td>0.860</td>
<td>0.403</td>
</tr>
<tr>
<td>WNN-GIP</td>
<td>0.846</td>
<td>0.323</td>
<td>0.813</td>
<td>0.263</td>
<td>0.888</td>
<td>0.403</td>
<td>0.864</td>
<td>0.497</td>
</tr>
<tr>
<td>Ours</td>
<td>0.849</td>
<td>0.432</td>
<td>0.817</td>
<td>0.370</td>
<td>0.888</td>
<td>0.422</td>
<td>0.882</td>
<td>0.521</td>
</tr>
</tbody>
</table>