Towards Drug Repositioning: A Unified Computational Framework for Integrating Multiple Aspects of Drug Similarity and Disease Similarity

Ping Zhang, Fei Wang, Jianying Hu

IBM T.J. Watson Research Center
Disclosure and learning objective

- Disclosure: All authors are employees of IBM
- Learning Objective: After participating in this activity the learner should be better able to:
  - Recognize the benefits of drug repositioning
  - Measure drug-drug and disease-disease similarities from multiple aspects
  - Formulate an approach to integrate biological and clinical data for drug repositioning
Drug repositioning (also known as Drug repurposing, Drug re-profiling, Therapeutic Switching and Drug re-tasking) is the application of known drugs and compounds to new indications (i.e., new diseases).

<table>
<thead>
<tr>
<th>Drug</th>
<th>Original indication</th>
<th>New indication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viagra</td>
<td>Hypertension</td>
<td>Erectile dysfunction</td>
</tr>
<tr>
<td>Wellbutrin</td>
<td>Depression</td>
<td>Smoking cessation</td>
</tr>
<tr>
<td>Thalidomide</td>
<td>Antiemetic</td>
<td>Multiple Myeloma</td>
</tr>
</tbody>
</table>

The repositioned drug has already passed a significant number of toxicity and other tests, its safety is known and the risk of failure for reasons of adverse toxicology are reduced.
Shorter timelines & less risk

Most of current methods only focus on one aspect of drug/disease activities.

Few methods consider both drug information and disease information.

Few methods can determine interpretable importance of different information sources during the prediction.

Multiple data sources of drug and disease information

Drug
- Chemical Structure
- Target Proteins
- Side-effect Keywords

Disease
- Phenotype/Symptom
- Ontology
- Disease Gene

Calculate drug/disease similarities

- weight loss
- impotence
- dizziness
- blurred vision
- ……
Examples of Drug Similarity Calculations

- For each drug pairs, we calculated three types of similarities based on chemical structures, target proteins, and side effects.

- Drug Similarity of Chemical Structures $D_{\text{chem}}$. We calculated the drug pairwise similarity based on a chemical structure fingerprint corresponding to the 881 chemical substructures defined in PubChem database. The pairwise chemical similarity between two drugs $d$ and $d'$ is computed as the Tanimoto coefficient of their chemical fingerprints:

$$D_{d,d'}^{\text{chem}} = \frac{h(d) \cdot h(d')}{|h(d)| + |h(d')| - h(d) \cdot h(d')}$$

- Drug Similarity of Target Proteins $D_{\text{target}}$. We collected all target proteins for each drug from DrugBank. Then we calculated the pairwise drug target similarity between drugs $d$ and $d'$ based on the average of sequence similarities of their target protein sets:

$$D_{d,d'}^{\text{target}} = \frac{1}{|P(d)||P(d')|} \sum_{i=1}^{|P(d)|} \sum_{j=1}^{|P(d')|} SW(P_i(d), P_j(d'))$$

- Drug Similarity of Side Effects $D_{\text{se}}$. We obtained side effect keywords from SIDER, an online database containing drug side effect information extracted from package inserts using text mining methods. The pairwise side effect similarity between two drugs $d$ and $d'$ is computed as the Tanimoto coefficient of their side effect profiles:

$$D_{d,d'}^{\text{se}} = \frac{e(d) \cdot e(d')}{|e(d)| + |e(d')| - e(d) \cdot e(d')}$$
Examples of Disease Similarity Calculations

- For each disease pairs, we calculated three types of similarities based on disease phenotypes, disease ontology, and disease genes.

- Disease Similarity of Phenotypes $S_{s's'}^{\text{pheno}}$. The disease phenotypic similarity was constructed by identifying similarity between the MeSH terms appearing in the medical description ("full text" and "clinical synopsis" fields) of diseases from OMIM database. The pairwise disease phenotype similarity between two diseases $s$ and $s'$ is computed as the cosine of the angle between their feature vectors:

$$
S_{s's'}^{\text{pheno}} = \frac{\sum_{i=1}^{K} m(s)_i m(s')_i}{\sqrt{\sum_{i=1}^{K} m^2(s)_i} \sqrt{\sum_{i=1}^{K} m^2(s')_i}}
$$

- Disease Similarity of Disease Ontology $S_{s's'}^{\text{do}}$. The Disease Ontology (DO) is an open source ontological description of human disease, organized from a clinical perspective of disease etiology and location. The semantic similarity of two diseases $s$ and $s'$ is defined as the information content of their lowest common ancestor by:

$$
S_{s's'}^{\text{do}} = -\log \min_{x \in C(s,s')} p_x
$$

- Disease Similarity of Disease Genes $S_{s's'}^{\text{gene}}$. Disease-causing aberrations in the normal function of a gene define that gene as a disease gene. We collected all disease genes for each disease from "phenotype-gene relationships" field from OMIM database. Then we calculated the pairwise disease similarity between diseases $s$ and $s'$ based on the average of sequence similarities of their disease gene sets:

$$
S_{s's'}^{\text{gene}} = \frac{1}{|G(s)||G(s')|} \sum_{i=1}^{|G(s)|} \sum_{j=1}^{|G(s')|} SW(G_i(s), G_j(s'))
$$
A graphical illustration of the main idea

And more drug similarity networks
......

And more disease similarity networks
......

© 2014 IBM Corporation
DDR Algorithm Flowchart (Input and Output)

- Drug chemical structure similarity network
- Drug target protein similarity network
- Drug side effect similarity network
- Disease phenotype similarity
- Disease ontology similarity
- Disease gene similarity

A unified computational framework for drug repositioning hypothesis generation

Outputs:
1. Predicted additional drug-disease associations
2. Interpretable importance of different information sources
3. Latent drug and disease groups as by-products
DDR as an optimization problem

Notations and symbols of the methodology

- $D_k$: The $k$-th drug similarity matrix
- $S_l$: The $l$-th disease similarity matrix
- $U$: Drug cluster assignment matrix
- $V$: Disease cluster assignment matrix
- $\Lambda$: Drug-disease cluster relationship matrix
- $R$: Observed drug-disease association matrix
- $\Theta$: Densified estimation of $R$
- $\omega$: Drug similarity weight vector
- $\pi$: Disease similarity weight vector

- We aim to analyze the drug-disease network by minimizing the following objective:
  \[ J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \]

- The reconstruction loss of observed drug-disease associations:
  \[ J_0 = \| \Theta - U \Lambda V^T \|_F^2 \]
  Similar Drugs/diseases (latent groups) have similar behaviors

- The reconstruction loss of drug similarities:
  \[ J_1 = \sum_{k=1}^{K_d} \omega_k \| D_k - UU^T \|_F^2 + \delta_1 \| \omega \|_2^2 \]
  Reconstruct integrated drug/disease networks

- The reconstruction loss of disease similarities:
  \[ J_2 = \sum_{l=1}^{K_s} \pi_l \| S_l - VV^T \|_F^2 + \delta_2 \| \pi \|_2^2 \]

- Putting everything together, we obtained the optimization problem to be resolved:
  \[ \min_{U, V, \Lambda, \Theta, \omega, \pi} J, \text{ subject to } U \geq 0, V \geq 0, \Lambda \geq 0, \omega \geq 0, \omega^T 1 = 1, \pi \geq 0, \pi^T 1 = 1, P_{\Omega}(\Theta) = P_{\Omega}(R) \]

© 2014 IBM Corporation
BCD approach for solving the non-convex optimization problem

- **Block Coordinate Descent (BCD) strategy:** The BCD approach works by solving the different groups of variables alternatively until convergence. At each iteration, it solves the optimization problem with respect to one group of variables with all other groups of variables fixed.

**Algorithm 1:** A BCD Approach for Solving Problem (11)

1. Initialize $\omega = (1/K_d) \mathbf{1} \in \mathbb{R}^{K_d \times 1}$, $\pi = (1/K_s) \mathbf{1} \in \mathbb{R}^{K_s \times 1}$

2. Initialize $U$ and $V$ by performing Symmetric Nonnegative Matrix Factorization on $\tilde{D} = \sum_{k=1}^{K_d} \omega_k D_k$ and

   $\tilde{S} = \sum_{l=1}^{K_s} \pi_l S_l$.

3. **while** Not Converge **do**

4. Solve $\Theta$ as described in section 2 (as a **constrained Euclidean projection**) \[ \text{closed-form solution} \]

5. Solve $\omega$ and $\pi$ as described in section 3 (as a **standard Euclidean projection onto a simplex**)

6. Solve $\Lambda$ as described in section 4 (as a **nonnegative quadratic optimization problem**) \[ \text{Solved by Projected Gradient Descent (PGD) method} \]

7. Solve $U$ as described in section 5 (as a **nonnegative quadratic optimization problem**)

8. Solve $V$ as described in section 6 (as a **nonnegative quadratic optimization problem**)

9. **end while**

Computational complexity is $O(Rrmn)$, where $R$ is the number of BCD iterations, and $r$ is the average PGD iterations when updating $\Lambda$, $U$, and $V$. © 2014 IBM Corporation
Data Description

- Benchmark dataset was extracted from NDF-RT, spanning 3,250 treatment associations between 799 drugs and 719 diseases
- **Three** $799 \times 799$ matrices were used to represent drug similarities between 799 drugs from different perspectives
- **Three** $719 \times 719$ matrices were used to represent disease similarities between 719 human diseases from different perspectives

![Graphs showing distribution of diseases per drug and drugs per disease](image)
The averaged ROC comparison of five drug repositioning approaches generated from 50 runs of 10-fold drug-based cross-validation.
Distribution of averaged weights and standard deviations of the similarity weight vectors obtained by DDR

(a) Drug similarity weight vector

(b) Disease similarity weight vector
Top 10 drugs for diseases Alzheimer's Disease (AD) and Systemic Lupus Erythematosus (SLE) based on leave-one-disease-out DDR predictions

(a) Top 10 drugs predicted for AD

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction Score</th>
<th>Clinical Evidence?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selegiline*</td>
<td>0.7091</td>
<td>—</td>
</tr>
<tr>
<td>Carbidopa</td>
<td>0.6924</td>
<td>No</td>
</tr>
<tr>
<td>Amantadine</td>
<td>0.6897</td>
<td>No</td>
</tr>
<tr>
<td>Procyclidine</td>
<td>0.6826</td>
<td>No</td>
</tr>
<tr>
<td>Valproic Acid*</td>
<td>0.6745</td>
<td>—</td>
</tr>
<tr>
<td>Metformin</td>
<td>0.6543</td>
<td>Yes</td>
</tr>
<tr>
<td>Bexarotene</td>
<td>0.6426</td>
<td>Yes</td>
</tr>
<tr>
<td>Neostigmine</td>
<td>0.6385</td>
<td>No</td>
</tr>
<tr>
<td>Galantamine*</td>
<td>0.6348</td>
<td>—</td>
</tr>
<tr>
<td>Nilvadipine</td>
<td>0.6159</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Repositioning candidates

(b) Top 10 drugs predicted for SLE

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction Score</th>
<th>Clinical Evidence?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desoximetasone</td>
<td>0.7409</td>
<td>No</td>
</tr>
<tr>
<td>Azathioprine*</td>
<td>0.7269</td>
<td>—</td>
</tr>
<tr>
<td>Leflunomide</td>
<td>0.7078</td>
<td>Yes</td>
</tr>
<tr>
<td>Fluorometholone</td>
<td>0.7054</td>
<td>No</td>
</tr>
<tr>
<td>Triamcinolone*</td>
<td>0.6862</td>
<td>—</td>
</tr>
<tr>
<td>Beclomethasone</td>
<td>0.6522</td>
<td>No</td>
</tr>
<tr>
<td>Etodolac</td>
<td>0.6445</td>
<td>No</td>
</tr>
<tr>
<td>Hydroxychloroquine*</td>
<td>0.6374</td>
<td>—</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>0.6371</td>
<td>Yes</td>
</tr>
<tr>
<td>Mercaptopurine</td>
<td>0.6150</td>
<td>No</td>
</tr>
</tbody>
</table>

* denotes the drug is known and approved to treat the disease
We proposed a general computational framework, to explore drug-disease associations from multiple drug/disease sources.

Our method could help generate drug repositioning hypotheses, which will benefit patients by offering more effective and safer treatments.

The computational framework and its solution can be used in other applications (gene-disease, drug-patient, etc.)
Thank you! | Questions?

Ping Zhang: pzhang@us.ibm.com
Fei Wang: fwang@us.ibm.com
Jianying Hu: jyhu@us.ibm.com