DRUG TARGET PREDICTION USING SEMANTIC LINKED DATA

Bin Chen
Ph.D. candidate
School of Informatics and Computing, Indiana University

binchen@indiana.edu
http://cheminfo.informatics.indiana.edu/~binchen

CSHALS, Feb 23, 2012
Beyond Data Integration

Chem2Bio2RDF.ORG
What is Drug Target?
Why Drug Target Prediction?
How to predict drug target?
Drug 1 bind Target 1

- Substructure
- Side effect
- Chemical ontology
- Gene expression profile

From Ligand (drug) perspective
From target perspective

Drug 1 bind Target 1

• Sequence
• 3D structure
• Gene Ontology
• Ligand

Target 2
Troglitazone

- bind to PPARG
- bind to ACSL4
- bind to PPARA

Rosiglitazone

- bind to PPARG

Pioglitazone

- bind to PPARG

Eicosapentaenoic Acid

- pathway to PPAR signaling pathway

Response to nutrient

- pathway to GO

hypoglycemic drug

- pathway to GO
Topology is important for association
Semantic is important for association

Cmpd 1 \(\text{bind}\) Protein 2 \(\text{bind}\) Cmpd 2 \(\text{bind}\) Protein 1

Cmpd 1 \(\text{bind}\) Protein 2 \(\text{hasGO}\) GO:00001 \(\text{hasGO}\) Protein 1

Cmpd 1 \(\text{bind}\) Protein 2 \(\text{PPI}\) Protein 1

Cmpd 1 \(\text{hasSideEffect}\) hypertension \(\text{hasSideEffect}\) Cmpd 2 \(\text{bind}\) Protein 1

Cmpd 1 \(\text{hasSubstructure}\) substructure 1 \(\text{hasSubstructure}\) Cmpd 2 \(\text{bind}\) Protein 1
SLAP
For Drug Target Prediction

(Semantic Link Association Prediction)
Path finding

>300k nodes, >1 million edges

Drug: Troglitazone

Target: PPARG
Let graph as $G(V, E)$, $P_l(s \rightarrow t)$ as the $l$th shortest path from node $s$ to $t$. $e_{i \rightarrow j}$ as the edge from node $i$ to node $j$. $R_{i,j}$ as the linked (relation) type of $e_{i,j}$.

We assume it has an equal probability surfing node $i$ to its neighbor node $j$ within the same type, thus:

$$p(e(i \rightarrow j)) = \frac{1}{\sum_{k=1}^{n} R_{i,n} = R_{i,j}}$$

where $k$ is the degree of node $i$. 
As the probability of each edge is independent, the probability from $s$ to $t$:

$$p(P_t(s \rightarrow t)) = p(P_t(e_1 \rightarrow e_2, e_2 \rightarrow e_3, ..., e_{m-1} \rightarrow e_m)) = \prod_{i=1}^{m-1} e_{i \rightarrow i+1}$$

where $m$ is the number of nodes in the path. Since $p$ is very small, the logarithm is applied,

$$\log(p(P_t(s \rightarrow t))) = \sum_{i=1}^{m-1} \log(e_{i \rightarrow i+1})$$

Accordingly, the probability from $t$ to $s$:

$$p(P_t(t \rightarrow s)) = p(P_t(e_m \rightarrow e_{m-1}, ..., e_3 \rightarrow e_2, e_2 \rightarrow e_1)) = \prod_{i=1}^{m-1} e_{i+1 \rightarrow i}$$

$$\log(p(P_t(t \rightarrow s))) = \sum_{i=1}^{m-1} \log(e_{i+1 \rightarrow i})$$

We consider the graph as undirected, then we take the average as the probability between $s$ and $t$:

$$p(P_t(s,t)) = (p(P_t(s \rightarrow t)) + p(P_t(t \rightarrow s)))/2$$
Path pattern

Path examples:

Path pattern:
Randomly sample 100,000 drug target pairs, yielding 453,087 paths, 35 patterns

- Plot Path pattern score distribution
- Convert path score to path z score
**Statistical Model---association score**

\[ \text{raw score}(s, t) = \sum_{l} n \frac{\log(p(P_l)) - \theta(\log(P_l))}{\sigma(\log(P_l))} \]

where \( \log(p(\log(P_l))) > \theta(\log(P_l)) \); \( n \) is the number of shortest paths between two objects \( s \) and \( t \); \( \theta(\log(P_l)) \) and \( \sigma(\log(P_l)) \) are expected mean, expected sd of the pattern to which \( P_l \) belongs.

Fit association scores of random pairs to normal distribution.
Association Score distribution among different pairs
Comparing SLAP with link prediction methods in social science

ROC curve

AUC=0.92
Polypharmacology profile comprises of association scores of one drug against over one thousand targets
Drug similarity network

• Nodes present drugs

• Two nodes are linked if they are similar in terms of biological function.

• Nodes are colored by their therapeutic indications
Dissimilar Drugs have same indication

Insomnia related drugs
Drug repurposing

Anti-Parkinson

Carbinoxamine

Cycrimine

allergic rhinitis

http://www.ebi.ac.uk/chebi/searchId.do?chebilstId=3398
Summary

- Semantic Link association can be assessed by topology and semantics of the network
- Domain knowledge plays an important role!
Team

- Prof. Ying Ding
- Prof. David Wild
http://chem2bio2rdf.org/slap
Thanks!
Backup slides
**SLAP Pipeline**

### (a) Raw Data Sets

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubChem</td>
<td>Chemotherapy Database</td>
</tr>
<tr>
<td>ChEBI</td>
<td>Chemicals and Biologicals Database</td>
</tr>
<tr>
<td>DrugBank</td>
<td>Drugs and Biologicals Database</td>
</tr>
<tr>
<td>UniProt</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>UniProtKB-GOA</td>
<td>Gene Ontology Database</td>
</tr>
<tr>
<td>HGNC</td>
<td>Human Genome Nomenclature Database</td>
</tr>
<tr>
<td>SIDER</td>
<td>Small Molecule Database</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>HPRD</td>
<td>Human Protein Reference Database</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>Chemicals and Biologicals Database</td>
</tr>
<tr>
<td>TTD</td>
<td>Therapeutic Targets Database</td>
</tr>
<tr>
<td>BindingDB</td>
<td>Binding Database</td>
</tr>
<tr>
<td>CTD</td>
<td>Chemicals and Therapeutics Database</td>
</tr>
<tr>
<td>PDSP</td>
<td>Pharmacological Database of Screening Profiles</td>
</tr>
</tbody>
</table>

### (b) Ontological level schema

- Gene ontology
- Pathway
- Chemical ontology
- Gene family
- Disease
- Tissue
- Target
- Chemical compound
- Substructure

### (c) Semantic Linked Data

- Integration
- Path finding
- Path filtering

### (d) Paths (length <4) between Two Nodes

- Drug: Troglitazone
- Target: PPARG

### (e) Significant Paths between Two Nodes

- Path filtering

### (f) Statistical Models

1. Edge weight:
   
   \[ p(c(i \rightarrow j)) = \frac{1}{\sum_{k=1}^{n} R_{i,k}} = R_{i,j} \]

2. Path score:
   
   \[ p(P(t \rightarrow s)) = p(P(c_{1},c_{2},...,c_{n})) = \prod_{i=1}^{n-1} c_{i+1-i} \]

3. Association score
   
   \[ \text{raw score}(s,t) = \sum_{i} \frac{\log(p(P_i)) - \theta(\log(P_i))}{\sigma(\log(P_i))} \]
<table>
<thead>
<tr>
<th>Path patterns</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical/Drug→bind→Target→bind→Chemical/Drug→bind→Target</td>
<td>0.850</td>
</tr>
<tr>
<td>Chemical/Drug→bind→Target→hasGo→GO→hasGO→Target</td>
<td>0.824</td>
</tr>
<tr>
<td>Chemical/Drug→hasSubstructure→SubStructure→hasSubstructure→Chemical/Drug→bind→Target</td>
<td>0.620</td>
</tr>
<tr>
<td>Chemical/Drug→express→Target→hasPathway→Pathway→hasPathway→Target</td>
<td>0.495</td>
</tr>
<tr>
<td>Chemical/Drug→express→Target→hasTissue→Tissue→hasTissue→Target</td>
<td>0.501</td>
</tr>
<tr>
<td>Chemical/Drug→express→Target→PPI→Target</td>
<td>0.501</td>
</tr>
</tbody>
</table>