11-925 Computational Biology Lab.
Drug Repositioning Project Report

Suyoun Kim
suyoun@cs.cmu.edu

Vision

The goal of the project is to build a system which is capable of identifying the new disease indication from known drugs and compounds information, and complete the literature review on the drug repositioning research area.

Motivation

Drugs interact with a protein target which is effective in disease of interest. Drugs may also interact with additional proteins that are not their primary therapeutic targets, resulting in unexpected side effects. These unexpected drug targets are usually unwanted and harmful; however, they can occasionally be beneficial and lead to different therapeutic indications. Since drug discovery and design is a very expensive and time-consuming process, there are many drugs whose target proteins (including the primary target and off-targets) have not yet been characterized. Thus*, in silic*o prediction of drug-target interactions plays an significant role in identifying and developing new uses of existing or abandoned drugs and reducing laboratory work and the cost of the experimental determination of drug-target interaction.

- **Who are the top researchers in the field?**

Yoshihiro Yamanishi
Division of System Cohort, Medical Institute of Bioregulation, Kyushu University
http://www.bioreg.kyushu-u.ac.jp/labo/systemcohort/yamanishi/index.html

Butte AJ
Department of Medicine, Stanford Center for Biomedical Informatics Research
http://buttelab.stanford.edu

Elena Marchiori
Associate professor, Radboud University in Nijmegen Netherlands
http://cs.ru.nl/~elenam/

Anil Goud Jegga
Assistant Professor, UC Department of Pediatrics
http://www.cincinnatichildrens.org/research/divisions/b/bmi/labs/aronow-jegga/default/

Robert Preissner
University Berlin
http://bioinformatics.charite.de/main/content/index.php

- **In silico Approach**

1) Focused on composition: chemical or molecular features of drugs
   a) network based nature of biological information to assess the possible new indication.
   b) quantitative chemical measures from drugs and targets

2) Based on knowledge about diseases: underlying processes or their symptomatology
   a) symptomatology, known treatments or pathological information
   b) “Guilt-by-association’ principle : two diseases are related when a similar treatment is

prescribed for both.

c) “Promiscuous” : side-effects information to relate drugs and diseases.

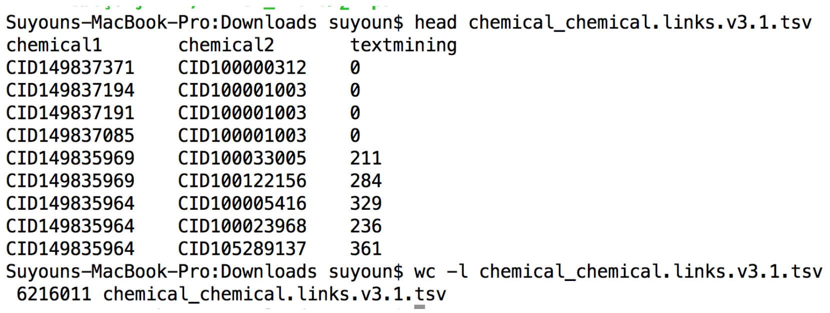
- **Where have they published?**

Nature
Briefing Bioinformatics (Impact Factor: 5.298, 5-Yr impact factor: 7.510)
Bioinformatics
Nucleic Acids Research
Pacific Symposium on Biocomputing

- **Source Databases**

Useful link: http://www.library.auckland.ac.nz/subject-guides/med/pharm.htm
Resources listed above link provide an overview of a drug, including aspects such as mode of action, pharmacodynamics, pharmacokinetics, therapeutic use(s), side effects, contraindications, targets etc. Resources include:
Drugdex; Martindales; AHFS (American Hospital Formulary Service);
MIMS; MedSafe Data Sheets; NZ Pharmaceutical Schedule;
NZ Universal List of Medicines (NZULM); NZ Formulary (NZF);
Australian Medicines Handbook; Australian Pharmaceutical Formulary & Handbook;
British National Formulary (BNF); Pharmaceutical Practice;
The Free Dictionary; RxList; DrugBank

STITCH 3.1 (Search Tool for Interactions of Chemicals)
http://stitch.embl.de/

<table>
<tr><td>Statistics</td><td>interactions for between 300,000 molecules 2.6 million proteins from 1133 organisms.</td></tr>
<tr><td>Flatfile</td><td>chem-chem 37.4 Mb

<pre>
Suyouns-MacBook-Pro:Downloads suyoun$ head chemical_chemical.links.v3.1.tsv
chemical1   chemical2   textmining
CID149837371   CID100000312   0
CID149837194   CID100001003   0
CID149837191   CID100001003   0
CID149837085   CID100001003   0
CID149835969   CID100033005   211
CID149835969   CID100122156   284
CID149835964   CID100005416   329
CID149835964   CID100023968   236
CID149835964   CID105289137   361
Suyouns-MacBook-Pro:Downloads suyoun$ wc -l chemical_chemical.links.v3.1.tsv
 6216011 chemical_chemical.links.v3.1.tsv
</pre>
chem-prot 1.5Gb</td></tr>
<tr><td>Identifier</td><td>CID123456789: number stands for the PubChem compound identifier
Prot: NNNNN.aaaaaa: NCBI taxonomy species identifier, RefSeq/Ensembl-identifier</td></tr>
</table>

SIDER 2 (Side Effect Resource) : March 16, 2012. information on marketed medicines and their recorded adverse drug reactions. extracted from public documents and package inserts. contains side effect frequency, drug and side effect classifications, links to further information, drug-target
http://sideeffects.embl.de/

Statistics	SE: 4192 drugs: 996 SE-drugs: 99,423 pairs with freq info: 40.8% drug, placebo # of drug side effect in different freq. ranges.
Flatfile	meddra_adverse_effect: STITCH id, concept id, drug, SE, concept type, MedDRA adverse_effect_raw, indications_raw_tsv: medical concepts label, concept id, SE meddra_freq_parsed: freq STITCH id, label, concept id, concept name, placebo?, freq[postmarketing, rare, infrequent, frequent, or exact %], lower bound, upper bound, MedDRA info

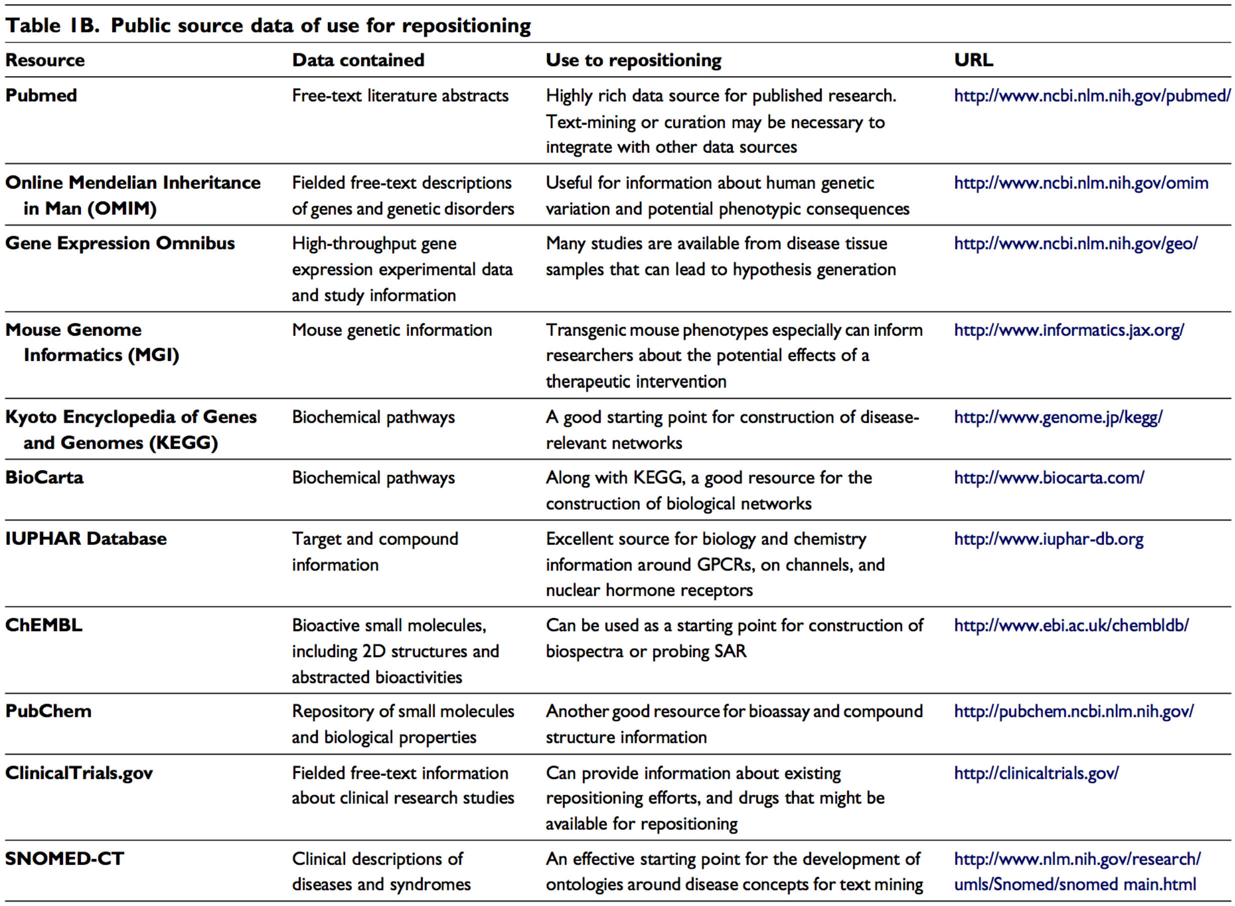


Table 1B. Public source data of use for repositioning

Resource	Data contained	Use to repositioning	URL
Pubmed	Free-text literature abstracts	Highly rich data source for published research. Text-mining or curation may be necessary to integrate with other data sources	http://www.ncbi.nlm.nih.gov/pubmed/
Online Mendelian Inheritance in Man (OMIM)	Fielded free-text descriptions of genes and genetic disorders	Useful for information about human genetic variation and potential phenotypic consequences	http://www.ncbi.nlm.nih.gov/omim
Gene Expression Omnibus	High-throughput gene expression experimental data and study information	Many studies are available from disease tissue samples that can lead to hypothesis generation	http://www.ncbi.nlm.nih.gov/geo/
Mouse Genome Informatics (MGI)	Mouse genetic information	Transgenic mouse phenotypes especially can inform researchers about the potential effects of a therapeutic intervention	http://www.informatics.jax.org/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Biochemical pathways	A good starting point for construction of disease-relevant networks	http://www.genome.jp/kegg/
BioCarta	Biochemical pathways	Along with KEGG, a good resource for the construction of biological networks	http://www.biocarta.com/
IUPHAR Database	Target and compound information	Excellent source for biology and chemistry information around GPCRs, on channels, and nuclear hormone receptors	http://www.iuphar-db.org
ChEMBL	Bioactive small molecules, including 2D structures and abstracted bioactivities	Can be used as a starting point for construction of biospectra or probing SAR	http://www.ebi.ac.uk/chembldb/
PubChem	Repository of small molecules and biological properties	Another good resource for bioassay and compound structure information	http://pubchem.ncbi.nlm.nih.gov/
ClinicalTrials.gov	Fielded free-text information about clinical research studies	Can provide information about existing repositioning efforts, and drugs that might be available for repositioning	http://clinicaltrials.gov/
SNOMED-CT	Clinical descriptions of diseases and syndromes	An effective starting point for the development of ontologies around disease concepts for text mining	http://www.nlm.nih.gov/research/umls/Snomed/snomed main.html

Literature Review

1) drug-target prediction

Alaimo, Salvatore, et al. "Drug-Target interaction prediction through Domain-Tuned Network Based Inference." Bioinformatics (2013).

- goal: predicting DTI (drug-target interaction)
- method: new network-based inference, called DT-hybrid: recommendation technique by domain-based knowledge including drug-target similarity, top-20 auc
- database: drugbank
- result: more reliable predicting drug-target interaction(DTI)

Wang, Yuhao, and Jianyang Zeng. "Predicting drug-target interactions using restricted Boltzmann machines." Bioinformatics 29.13 (2013): i126-i134.

- They predict DTI and drug modes of action by using two layer graphical mode(=restrict boltzmann machine = RBM) which specifying drugs types separately and reveals that it performs better than simply mixing types

- goal: predicting DTI or drug modes of action
- method: two layer graphical model, called restricted Boltzmann machine, integrating multiple types of DTIs. 1) hidden units (binary) 2) visible units observed types [Direct?, Indirect?, Binding?, Inhibition?, ...]
- database: STITCH, provides mode of action for interactions between proteins and chemicals
- result: improved AUC 89.6

Takarabe, Masataka, et al. "Drug target prediction using adverse event report systems: a pharmacogenomic approach." Bioinformatics 28.18 (2012): i611-i618.

- They first define the pharmacological similarity and apply it in pairwise kernel regression for predicting DTI. They use the similarity between two kernels: drug kernel func. and target kernel func.
- goal: predicting DTI
- method: define pharmacological similarity of drugs and genome sequence similarity of target proteins
- result: especially effective for predicting DTI that could not be expected from drug chemical structures. 1874 drugs known targets, 2519 drugs without known targets

2) drug-drug similarity based on analysis of their side effects

Kuhn, Michael, et al. "A side effect resource to capture phenotypic effects of drugs." *Molecular systems biology* 6.1 (2010).

- contributes: SIDER with number of analysis. make dictionary by text mining from PDF->text & SPL(kind of XML) UMLS(United Medical Language System), COSTART(Coding Symbols for a Thesaurus of Adverse Reaction Terms). Mapped drug names to PubChem identifiers.
- datasets: only used public resource, FDA-approved drugs
- side-effect: phenotypic responses of the human organism to drug treatment
- common side-effect: shared underlying mechanisms of action.
- 1) predicting the possible side effects of drug candidates
- 2) predicting novel drug-target interactions
- 3) drug re-purposing Campillos , 2008
- Result1: 70% of drugs have 10~100 side-effects
- Result2: 55% of side-effect only occurs in below 10 drugs
- Result3: 25% of side-effect only occurs in one drug class
- Result4: 12/14 drug classes over-represented side-effects.
- got 500/888 drugs’ frequency.
- Result5: 14% of side-effect only more frequent placebo than side effect

Campillos, Monica, et al. "Drug target identification using side-effect similarity."*Science* 321.5886 (2008): 263-266.

- drug target: so far molecular or cellular feat. => phenotypic side-effect similarity
- database: 746 drugs, 1018 side effects, drug-drug relation, 261 formed by dissimilar drugs from different therapeutic indications.

- evaluation: 20 of drug-drug relation, validate 13 drug-target relation
- results: 11/20 reveal inhibition constants =< 10 micromolar, 9/20 were tested and confirmed in cell assays
- method: i-side-effect: 1) rareness score r_i=-log(|I|/855) 2) correlation score c_i 3) individual correlation score c_i hierarchical clustering and weighting algorithm 4) raw score: each drug pair d, □ the raw side-effect similarity score = sum(rareness*correlation) for all side-effect they share.
- 3215 unique high scoring drug pairs
- ATC(Anatomical Therapeutic Chemical) Classification System 141, 3-4 digit code. calculate correlation between ATC categories.
- previous work: 1) similar protein binding profiles => similar side effects, 2) phenotypic assays => chemical similarity measure, 3) docking strategies
- result1: small overlap in (side-effect similarity) and (chemical similarity) [50% prob]
- result2: 956/2903(known shared target) 1947/2903(novel drug pairs) [754 unexpected shared target]
- result3: predicted drugs to proteins associated with different therapeutic categories.
- validate: vitro and cell assays

Smith, Richard B. "Repositioned drugs: integrating intellectual property and regulatory strategies." *Drug Discovery Today: Therapeutic Strategies* 8.3 (2012): 131-137.

- biological data mining:

Chiang and Butte: identify all disease-drug combinations.

- database: SNOMED-CD, DRUGDX (FDA approved uses of drugs and physician prescribed off-label uses),
- result: uncover novel 5500 pairs
- chemical data mining:
- text based mining of published knowledge

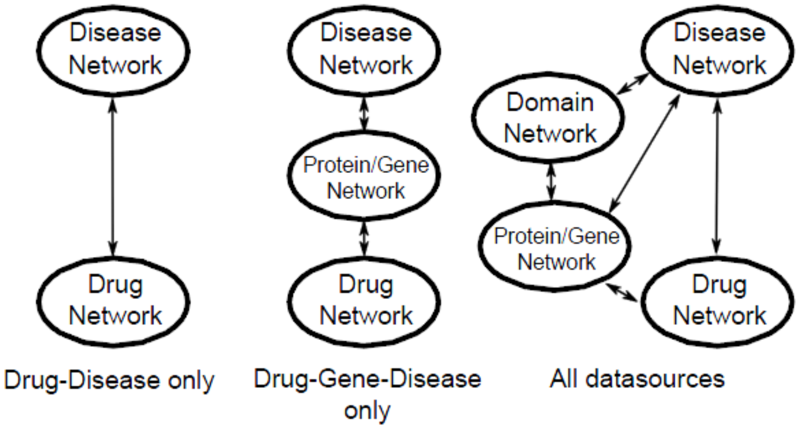
Loging, William, et al. "Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing." *Drug Discovery Today: Therapeutic Strategies* 8.3 (2012): 109-116.

3) network analysis also in combination with side effects

Martınez, Vıctor, et al. "Network-based drug-disease relation prioritization using ProphNet."

- They used network-based method, which characterize the degree of relation between drugs sets and diseases sets, for determining the best data source for drug repositioning.
- method: new method based on network prioritization. Integrate = [symptomatology-based + molecular-based(expression or pathway)], ProphNet: network based prioritization tool.
    - each network□ gene-gene or drug interaction, node v□ drug or disease, weighted arcs□ interaction pair of nodes,
    - Two networks: 1) from same domain 2) from two different domains. using Adjacency matrices. (network connection)
    - finally build Global Graph, measure the degree of relation bet. two sets of nodes. Query

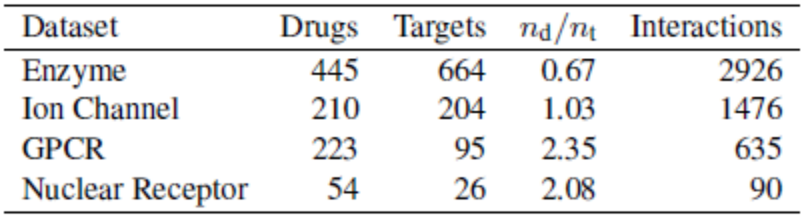
    Set and Target Set, path of networks 1) “propagation within a network” 2) “network-to-network propagation”
    - mean ranking position of the query disease 5080
- result: reveals that selection of the data sources to be integrated can be critical step.
- only drug-disease best.
- validate: LOO leave-one-out test to determine the best data sources for the drug repositioning task. ROC curves: TP: the rank of case disease is below the threshold.
- database: only consider drug-disease has clinical trials from ClinicalTrials.gov
    - OMIM for disease network.
    - DrugBank for drug network. 1109 drugs, 10906 interactions.
    - DOMINE and InterDom for protein domain network: 5490 domain, 48778 interactions.
    - Pfam and UniProt for domain-drug
    - HPRD for protein network, 8919 protein 64662 interaction
    - DrugBank for drug-protein 2860
    - OMIM for gene-disease 1393
    - Pfam domain-domain
    - DailyMed 1337 drug-disease



van Laarhoven, Twan, Sander B. Nabuurs, and Elena Marchiori. "Gaussian interaction profile kernels for predicting drug–target interaction." Bioinformatics 27.21 (2011): 3036-3043.

- goal: predicting DTI (drug-target interaction)
- method: interaction profiles of drugs in a network(binary vector which specifying present or absence of interaction with every target), then define a kernel on these profiles, Gaussian Interaction Profile (GIP) kernel. used regularized least square (RLS)
- result: improved AUC.
- Hypothesis is that drugs(targets) have similar pattern of interaction and non-interaction with the targets(drugs) of a drug-target interaction network,
- and are likely to show similar interaction behavior with respect to new targets.
- -. Method: They use interaction profiles of drugs and targets in a network,
- which are binary vectors specifying the presence or absence of interaction with every target in that network.
- -. Then, (they define a kernel on these profiles), and use Regularized Least Squares (RLS), for prediction DTI

- -. Evaluation: They tested and compare with methods from previous studies and (also use kernels based on chemical structure similarity and genomic sequence similarity information.)



Dataset	Drugs	Targets	$n_{\mathrm{d}}/n_{\mathrm{t}}$	Interactions
Enzyme	445	664	0.67	2926
Ion Channel	210	204	1.03	1476
GPCR	223	95	2.35	635
Nuclear Receptor	54	26	2.08	90

Mei, Jian-Ping, et al. "Drug–target interaction prediction by learning from local information and neighbors." Bioinformatics 29.2 (2013): 238-245.

- In order to deal with insufficient training data and the problem in current method Bipartite Local Model, which is not able to predict new candidates, they obtain the training data from their neighbors specifying the high chemical similarity.
- goal: predicting drug and four categories of target proteins.
- method: neighbor-based interaction-profile inferring (NII), then integrate into existing BLM method => BLM-NII method

Mizutani, Sayaka, et al. "Relating drug–protein interaction network with drug side effects." Bioinformatics 28.18 (2012): i522-i528.

- They predict side-effect by using protein binding profiles with SCCA-based approach, and they reveals that it better performs than chemical structure based method.
- goal: analysis to extract correlated sets of targeted proteins and side-effects.
- method: based on co-occurrence of drugs in protein-binding profile and side-effect profile,
- result: most of correlated sets were enriched with proteins that are involved in the same biological pathways, even if their molecular functions are different.
- database: KEGG, GO

Gottlieb, Assaf, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine." *Molecular systems biology* 7.1 (2011).

- goal: predict new drug indication.
- method: utilize multiple drug-drug and disease-disease similarity measure for the prediction task. For drugs, they used chemical based, side effect based, sequence based, closeness in a PPI network, GO based features, and for disease, phenotype based, semantic phenotype similarity, genetic based, signature based features are used.
- result: AUC=0.9
- validation: overlap with drug indication that are currently under clinical trials

In recent years, a variety of *in silico* methods have been developed to predict drug-target (Bleakley and Yamanishi, 2009; Faulon et al., 2008; Jacob and Vert,2008; Keiser et all., 2009; van Laarhoven et al., 2011; Yamanishi et all. Traditional methods rely on ligand-based approaches. The underlying idea is that similar drug components are likely to interact with similar proteins. These predictions are performed based on chemical structures of drug compounds, and protein sequences of targets and the currently known drug-protein interactions.

Another promising approach is to use pharmacological information such as drug side effects and adverse drug reactions. The use of side effect similarity has been recently proposed to infer whether two drugs share a target (Campillos et al., 2008; Yamanishi et al., 2010). Campillos et al.(2008) newly identified DTIs using side effect similarity. These methods are useful when chemical structures and side effects are correlated with each other to some extent.

One of the state-of-the-art approach is network-based method. Most of these methods rely on the idea of constructing bipartite network, were applied to predict drug-target interactions and infer repositioning candidates like Bleakley and Yamanishi, 2009; van Laarhoven et al., 2011. By using kernel function, multiple sources of information can be easily incorporated for high accuracy prediction.

Van Laarhoven et al.(2011) use the regularized least squares algorithm which is trained using a kernel that summarizes the information in the network. The author developed various kernels by taking into account chemical and genomic information.

In Victor Martinez (2013), new network-based prioritization method, ProphNet, was applied to prioritize drugs related to a query disease. Their method integrates a set of networks, (e.g. one network modeling gene-gene interactions, one for drug interactions), into a global network in which entities of different types are interconnected.

In this work, we propose a method for drug repositioning based on networks prioritization. It extends the ProphNet algorithm proposed in Victor Martinez (2013) by adding pharmacological knowledge. Side effect similarity among drugs is plugged into the model. To demonstrate the performance of the method, we conducted a wide experimental analysis with different possible configurations of data sources.

Approach to Implement

1. **Target Approach**
   Martınez, Vıctor, et al. "Network-based drug-disease relation prioritization using ProphNet."
   Paper: http://iwbbio.ugr.es/papers/iwbbio_102.pdf
   Website: http://genome2.ugr.es/prophnet/

2. **Goal**
   I have tried to tackle the problem by working on two subproblems.
   - Implement ProphNet to reproduce the results with our database
   - Apply side effect similarity

3. **Research Question**
   The hypothesis is whether adding side effect similarity information to ProphNet algorithm can be improved to predict the new drug disease interaction.

4. **Risk**
   - Domain Knowledge: I did not have a background with drug repositioning and did not know about the drug and disease databases.
   - Lack of Gold Standard Data: In order to evaluate the performance of our system fairly,

the drug disease interaction dataset that used in previous work requires but it is not available.

- Limitation of the ProphNet source code and document: The supplementary of this paper and the databases were not provided entirely, so I took time in understanding the algorithm fully.

Methods

Prioritize drugs and diseases

I have applied ProphNet to prioritize drugs and diseases. ProphNet algorithm proposed in previous works Victor Martinez (2013). To perform the prioritization task, first we define and build the data networks. The algorithm uses a graph representation of data sources where each node corresponds to a biological entity of a type of interest, for example, gene, protein, disease, protein domain, etc. The arcs between two nodes are labeled with a weight (from 0 to 1) and nodes in networks are connected by this weighted arcs. This arc represents an interaction or relationship between the pair of nodes and the weight of arcs represents the strength of the relationship.

The method integrates a set of networks, and there are two types of networks: one connecting entities of one type (same domain), the other network which entities of different types (different domain), i.e. the query and target sets belong to two different networks. These network connections are represented as adjacency matrices and each matrix is normalized.

ProphNet Algorithm:

The nodes of the networks adjacent to the target network to take a value based on their degree of relationship to the query set.

1. To measure the score(degree) of the path between the query entity and the target entity, the initial value is set to 1/|X| by default, where |X| is the cardinal of the query set and the target set
2. After the initial values have been set, these scores are propagated within each subnetwork and between networks. The score is propagated until convergence, taking into account the weight of the arc. The propagation of the score between networks can be calculated by assigning each node in the next subnetwork(B) directly connected to nodes in the previous subnetwork(A) the average of the values of the neighbour nodes in the previous subnetwork.(A)

The propagation of the score within each subnetwork:

$$x_{i+1} = (1-\alpha) * M * x_i + \alpha * x_0,$$

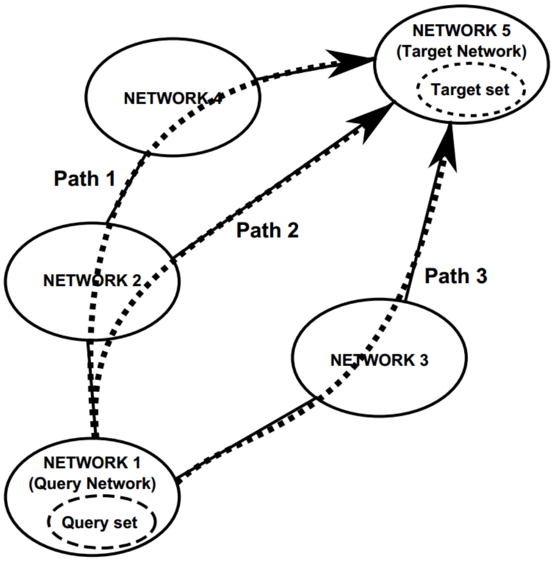
where alpha is a parameter which determines the importance of the prior information in the network, (0.9), M is the normalized adjacency matrix of the network, and x_i is a network node value.

The propagation of the score between subnetworks:

$$\Psi(v) = \frac{\sum^{x\in neig(v)} \Psi(x)}{|neig(v)|},$$

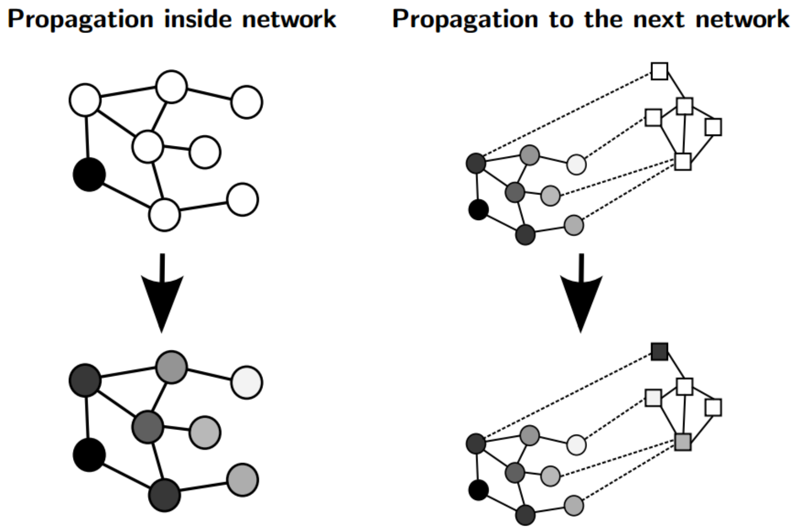
where neig(v) is the set of nodes from the current network which are connected with node v in the next network.

3. Finally, to calculate the score of relationship between the query and the target, the correlation between the score of target nodes and the score of the nodes from adjacent networks directly connected to the target nodes.



Pseudo code

```
Query set values are propagated inside Query Network.
paths ← Compute all paths from Query to Target Net.
for each node e in Target Network do
    Set node e as Target Set and propagate values inside Target Network
    for each path in paths do
        for each step in path do
            Propagate from current network to next network
            Propagate inside next network
        end for
    end for
    Se ← Correlate paths with Target Network values
end for
Sort Sx decrementally to obtain prioritized list
```



Side-effect similarity

We used as a data set 996 different FDA approved drugs and 4,192 different associated side effects, and a total of 99,423 drug-side effect pairs, all downloaded from the database SIDER (Side Effect Resource). From this data set, we calculate side effect similarity scores between drug pairs by using the method based on a rareness and correlation for each side effect proposed in Campillos et al.(2008).

Formally we are given sets, N = {996 drugs d} of drugs, D = {The set of all side effects associated with drug d}, S = {The set of all 4,192 side effects i}, and I = {The set of all drugs with side effect i}. The rareness score r_i is a reflection of the relative rareness of the side effect i. The rareness score is the negative logarithm of the fraction of the number of drugs with side effect i out of the total 996 drugs in the data set:

$$r_i = -\log_{10}\frac{|I|}{|N|}$$

The correlation score c_i for each side effect i can be represented as a distance matrix. First, The correlation between side effects I and j is:

$$\mathrm{Corr}(i,j) = \frac{|I\cap J|}{|I\cup J|}$$

The correlation between every possible pair of side effects can be calculated by the number of drugs that have both side effects I and j, divided by the number of drugs that have either side effect I or j (or both). Thus, any pairwise correlation between two side effects is between 0 and 1.

Then, individual correlation scores c_i were calculated for each side effect by hierarchical clustering of the above distance matrix proposed in “Volume Changes in Protein Evolution” by Gerstein, et al., which gave an individual correlation score to each side effect corresponding to its position in the dendrogram created by clustering analysis.

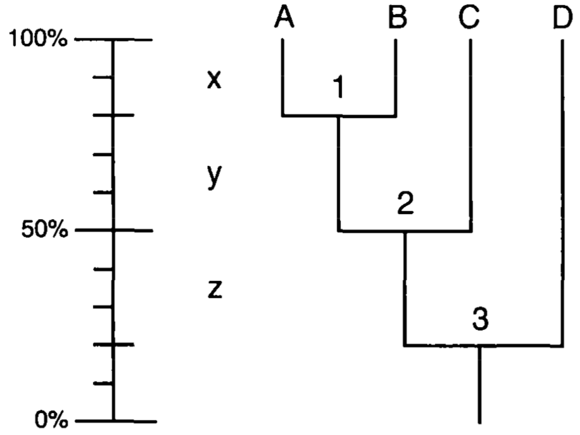
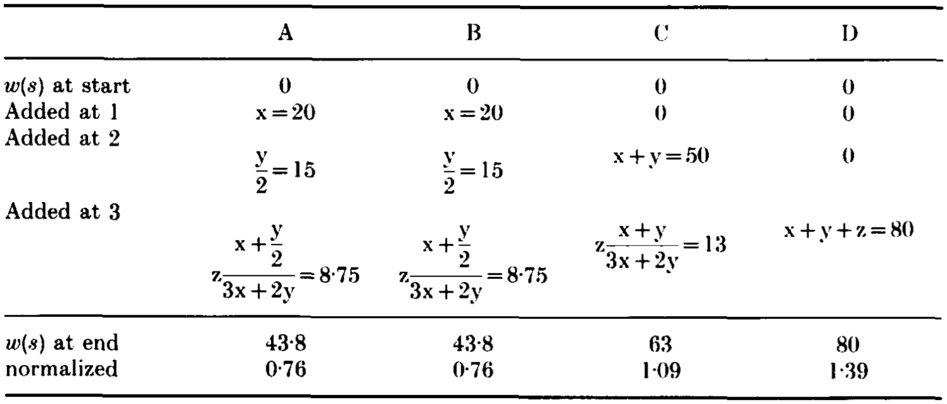


Figure 4. A worked example of our weighting method. The Figure shows a bifurcating tree with 4 sequences; A, B, C and D. A and B are 80% identical; the average identity between C and A or B is 50%; and the average identity between D and A, B, or C is 20%. The weights for each sequence, denoted $w(s)$, are calculated by visiting the nodes sequentially (first node 1, then 2, and finally 3), and adding increments to the total weight at each node. At the end the final weights are normalized, so that the average weight is 1. The calculation is summarized below:



	A	B	C	D
$w(s)$ at start	0	0	0	0
Added at 1	$x=20$	$x=20$	0	0
Added at 2	$\frac{y}{2}=15$	$\frac{y}{2}=15$	$x+y=50$	0
Added at 3	$z\frac{x+\frac{y}{2}}{3x+2y}=8{\cdot}75$	$z\frac{x+\frac{y}{2}}{3x+2y}=8{\cdot}75$	$z\frac{x+y}{3x+2y}=13$	$x+y+z=80$
$w(s)$ at end	43·8	43·8	63	80
normalized	0·76	0·76	1·09	1·39

A worked example for calculating the individual correlation score c_i is shown above. All sequences initially have a weight of 0. Then, we traverse the tree by visiting each node, going from 100% (leaves) to 0% (root), and for that node determine the weight increment for the left and right subtrees. The left subtree increment applies to all sequences in this subtree. and likewise for right subtree increment. The weight added to a subtree is the length of the edge connecting it to the node currently being visited. The length of this edge is measured from the last previously visited node of the subtree. It is apportioned between weights are updated according to the following formula:

$$w(s,b) \square w(s,b) + D(b) F(s,b).$$

where b is L or R for the left or right subtree, *s* runs over all sequences in a subtree. *D(b)* is the edge length to be apportioned. *w(s,b)* is the current weight of sequence *s* in subtree *t*. and *F(s,b)* is the weight fraction of sequence *s* in the subtree *t.*

$$F(s,b) = \{1, \square\square\ \square(\square,\square) = 0, \square\square\square\square\square\square\square\square\square, \square(\square,\square)/\textstyle\sum_{\square}\ \square(\square,\square)\}$$

Finally, a raw score was calculated for each drug pair d, e as follows:

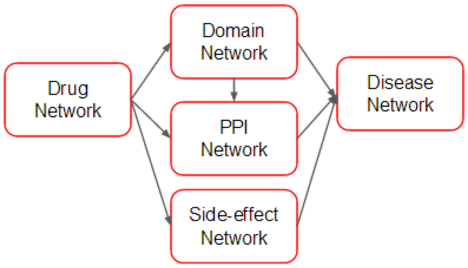
$$\forall d, e \in N, d \neq e,\ \textbf{Raw Side effect similarity}(d,e) = \sum_{i \in D \cap E} (r_i)(c_i)$$

That is, for any pair of different drugs, the raw side effect similarity score between them is the sum of the products of the rareness and correlation scores for all side effects they share.

System Overview

Design

The following diagram gives an overview of our system. Our system takes list of drugs as input and generate the prioritized list of disease. By using the disease drugs relation, the drug-drug relation, disease-disease relation, protein protein interaction, domain relation, and side-effect database as resources it performs the task of predicting and prioritizing drugs related to a query disease. The following diagram shows the pipeline of our system.



- **What is the challenge?**

To compare the performance between ProphNet work and our work fairly, I tried to use same test sets, however, ProphNet use different disease ID. They use OMIM disease ID, however, our system database uses KEGG disease ID. So I had to implement convertor to KEGG disease ID from OMIM disease ID to extract the gold standard test sets. These make the reproducing results be hard.

- **How to do it?**

As a solution, I downloaded KEGG disease database from the link and extract the pair of disease ID and disease name. ftp://ftp.genome.jp/pub/kegg/medicus/

Because the name of disease are not matched exactly from OMIM disease name and KEGG disease name, so I had to decide the criteria to convert. For matching the OMIM disease name to KEGG disease name, I converted if the two names are more than 80% identical. I used difflib.SequenceMatcher library. Finally, I could obtain the 321 drug-disease pairs.

Databases

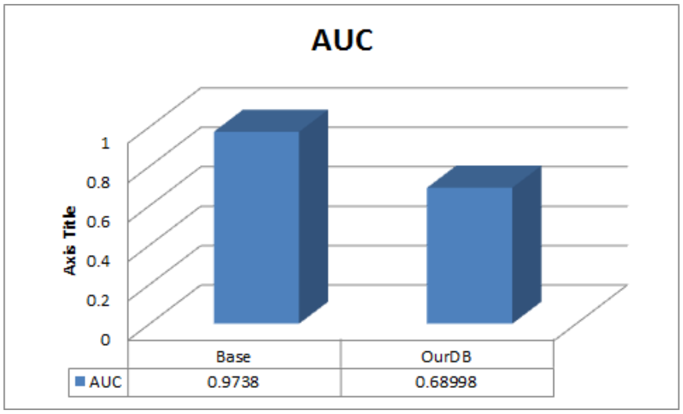
OMIM for disease network.
DrugBank for drug network. 1109 drugs, 10906 interactions.
DOMINE and InterDom for protein domain network: 5490 domain, 48778 interactions.
Pfam and UniProt for domain-drug
HPRD for protein network, 8919 protein 64662 interaction
DrugBank for drug-protein 2860
OMIM for gene-disease 1393
Pfam domain-domain

	Number of Entities	Number of Interactions	Source
Disease	5080	39458	OMIM
Domain	5490	48778	DOMINE InterDom
Drug	1109	10906	DrugBank
PPI	8919	64662	HPRD
Domain-Drug	-	1614	Pfam UniProt
Drug-Domain	-	2860	DrugBank
Protein-Disease	-	1393	OMIM

Results

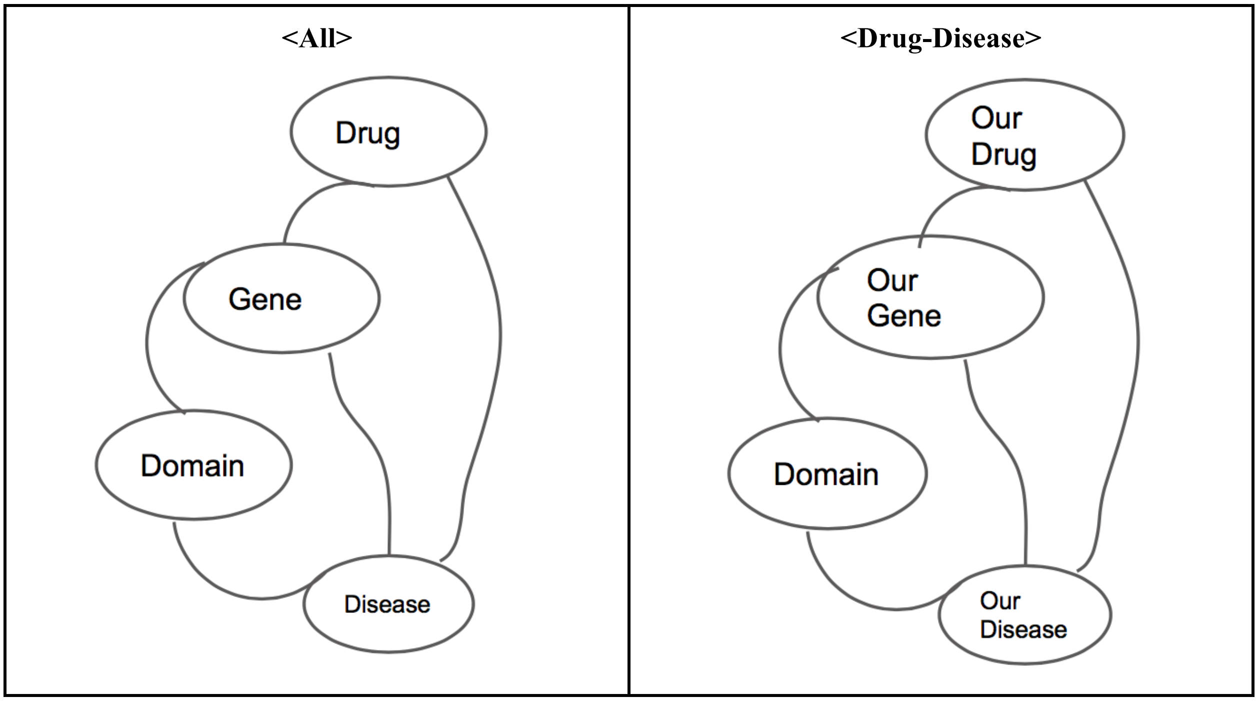
LOO test with our databases.

The drug-disease network has been computed by mapping disease names to UMLS concepts and matching these with drugs indications from DailyMed as described by (Gottlieb, 2011) For validation the networks, leave-one-out (LOO) test was performed for the global graph obtaining its performance by removing one known drug-disease relation, taking the drug as query set and checking the resultant disease ranking to measure performance. A ROC curve is created by plotting the fraction of true positives out of the positives vs. the fraction of false positives out of the negatives at various threshold settings. A true positive occurs when the rank of the case disease is below the threshold. A false positive occurs when a disease that is not in the case is ranked below the threshold. The area under the ROC curve (AUC value) was also computed to quantify gains.



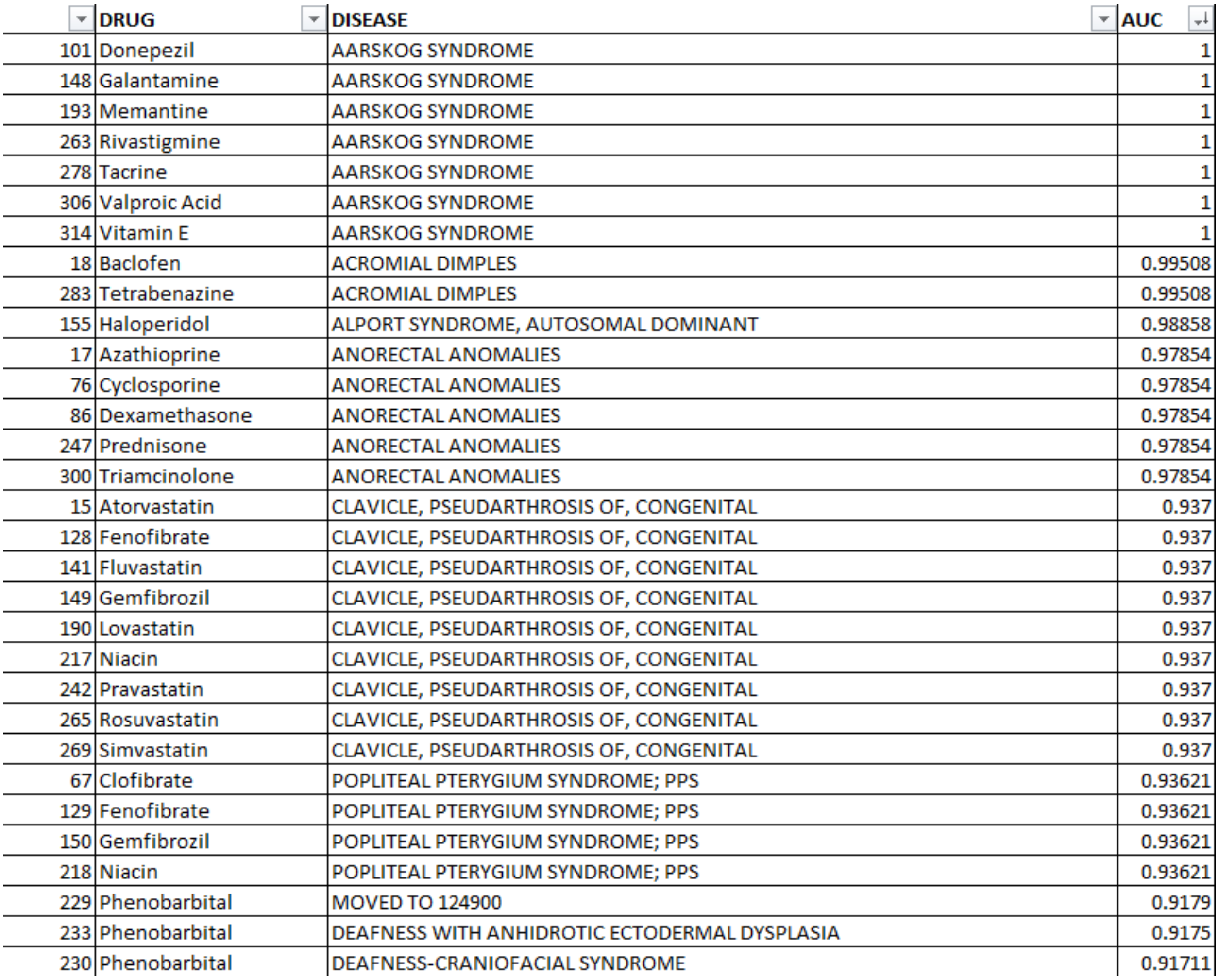
LOO test with test_drugs, and test_disease. with drug-disease path. The test set contains 321 drug-disease pairs. Tests with disease-drug networks obtained a 0.68998 AUC value and 803.2486± 1575.5888 mean ranking.

<network configuration>



	Number of Entities	Source
Disease	1066	KEGG
Drug	6792	DrugBank

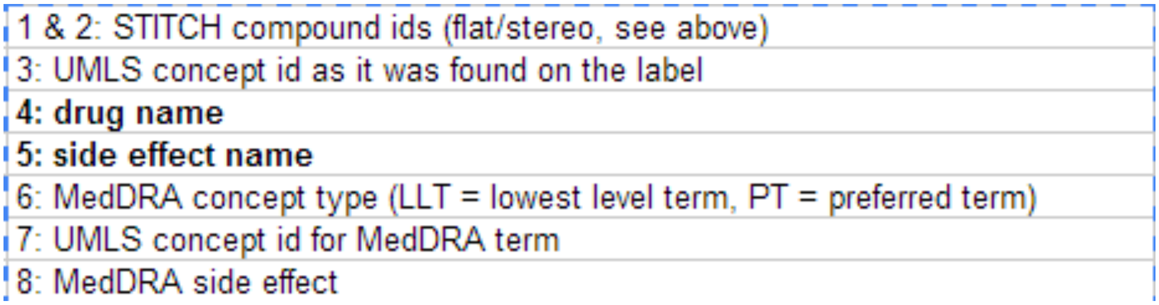
PPI	43759	BioGRID MINT



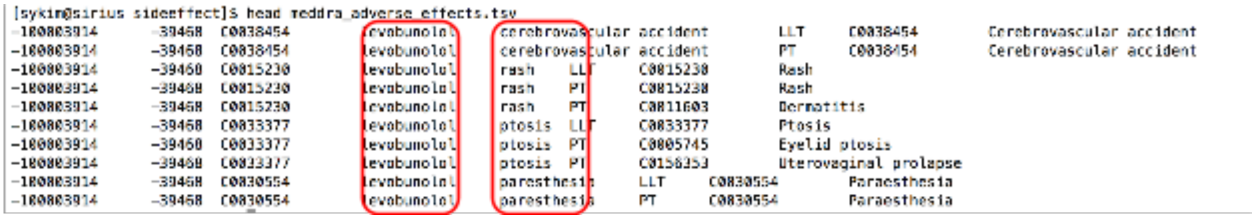
	DRUG	DISEASE	AUC
101	Donepezil	AARSKOG SYNDROME	1
148	Galantamine	AARSKOG SYNDROME	1
193	Memantine	AARSKOG SYNDROME	1
263	Rivastigmine	AARSKOG SYNDROME	1
278	Tacrine	AARSKOG SYNDROME	1
306	Valproic Acid	AARSKOG SYNDROME	1
314	Vitamin E	AARSKOG SYNDROME	1
18	Baclofen	ACROMIAL DIMPLES	0.99508
283	Tetrabenazine	ACROMIAL DIMPLES	0.99508
155	Haloperidol	ALPORT SYNDROME, AUTOSOMAL DOMINANT	0.98858
17	Azathioprine	ANORECTAL ANOMALIES	0.97854
76	Cyclosporine	ANORECTAL ANOMALIES	0.97854
86	Dexamethasone	ANORECTAL ANOMALIES	0.97854
247	Prednisone	ANORECTAL ANOMALIES	0.97854
300	Triamcinolone	ANORECTAL ANOMALIES	0.97854
15	Atorvastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
128	Fenofibrate	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
141	Fluvastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
149	Gemfibrozil	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
190	Lovastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
217	Niacin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
242	Pravastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
265	Rosuvastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
269	Simvastatin	CLAVICLE, PSEUDARTHROSIS OF, CONGENITAL	0.937
67	Clofibrate	POPLITEAL PTERYGIUM SYNDROME; PPS	0.93621
129	Fenofibrate	POPLITEAL PTERYGIUM SYNDROME; PPS	0.93621
150	Gemfibrozil	POPLITEAL PTERYGIUM SYNDROME; PPS	0.93621
218	Niacin	POPLITEAL PTERYGIUM SYNDROME; PPS	0.93621
229	Phenobarbital	MOVED TO 124900	0.9179
233	Phenobarbital	DEAFNESS WITH ANHIDROTIC ECTODERMAL DYSPLASIA	0.9175
230	Phenobarbital	DEAFNESS-CRANIOFACIAL SYNDROME	0.91711

Results with Side-effect similarity

We downloaded the f SIDER 2, which has been released on March 16, 2012. And I extracted the drug name and side effect name as I described the above method section.



1 & 2: STITCH compound ids (flat/stereo, see above)
3: UMLS concept id as it was found on the label
4: drug name
5: side effect name
6: MedDRA concept type (LLT = lowest level term, PT = preferred term)
7: UMLS concept id for MedDRA term
8: MedDRA side effect

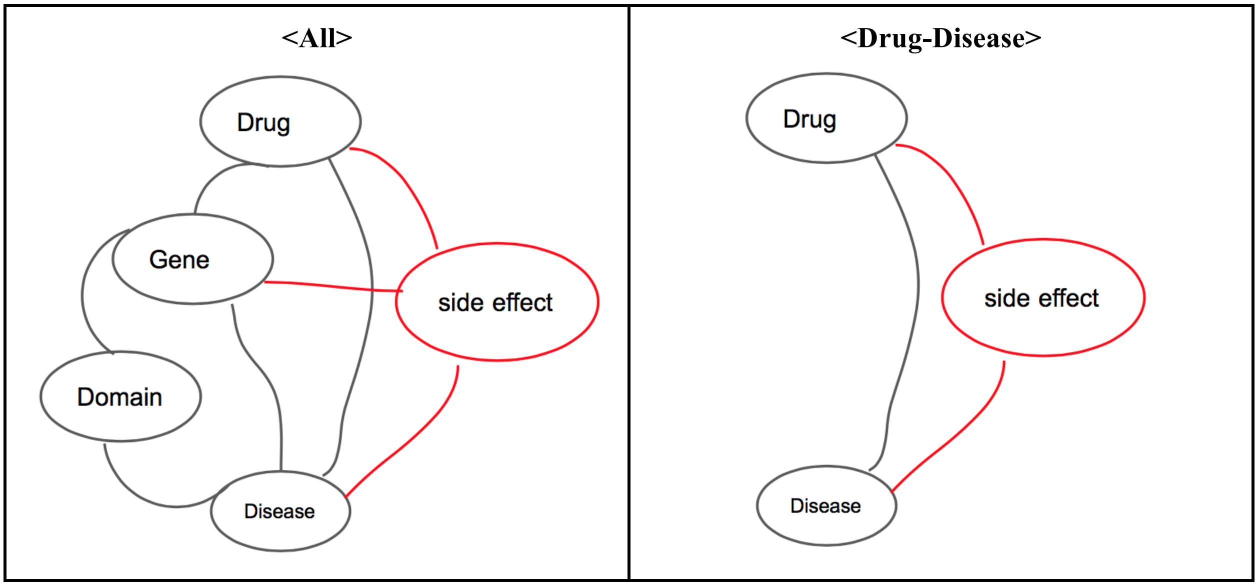


```
[sykim@sirius sideeffect]$ head meddra_adverse_effects.tsv
-100003914      -39468  C0038454        levobunolol     cerebrovascular accident        LLT     C0038454        Cerebrovascular accident
-100003914      -39468  C0038454        levobunolol     cerebrovascular accident        PT      C0038454        Cerebrovascular accident
-100003914      -39468  C0015230        levobunolol     rash    LLT     C0015230        Rash
-100003914      -39468  C0015230        levobunolol     rash    PT      C0015230        Rash
-100003914      -39468  C0015230        levobunolol     rash    PT      C0011603        Dermatitis
-100003914      -39468  C0033377        levobunolol     ptosis  LLT     C0033377        Ptosis
-100003914      -39468  C0033377        levobunolol     ptosis  PT      C0005745        Eyelid ptosis
-100003914      -39468  C0033377        levobunolol     ptosis  PT      C0158353        Uterovaginal prolapse
-100003914      -39468  C0030554        levobunolol     paresthesia     LLT     C0030554        Paraesthesia
-100003914      -39468  C0030554        levobunolol     paresthesia     PT      C0030554        Paraesthesia
```

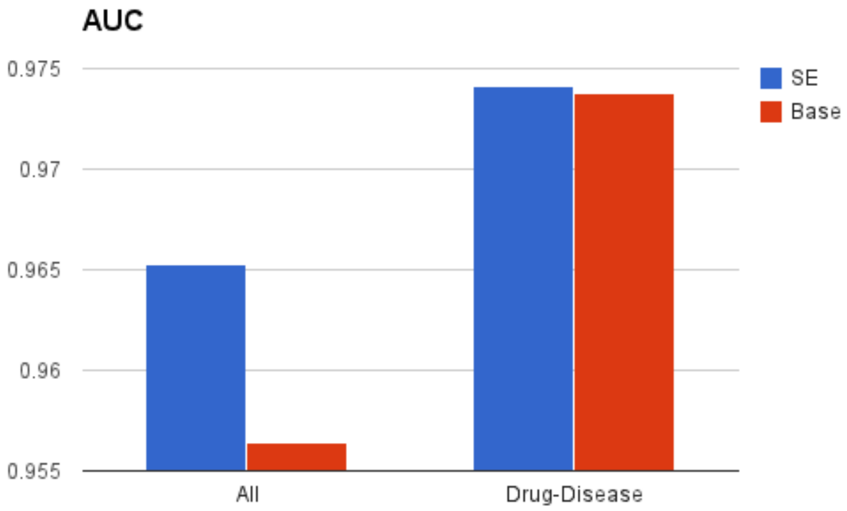
The number of side effect is 4192 and the number of drugs is 996, and 99,423 interactions are obtained.

# of SE	# of drugs	# of pairs
4192	996	99423

<network configuration>



	All		Drug-Disease	
	SE	**Base**	**SE**	**Base**
AUC	**0.96522**	0.95638	**0.97411**	0.9738
Mean ranking	**177.6627**	222.5318	**132.5183**	134
Std.Dev	**513.8512**	566.0316	**435.4639**	438



LOO tests consists of 1337 test cases (one for each explicit drug-disease relationship in the global graph). I measured the accuracy of the ranking for two different data configurations. The tests with all the data sources without the side-effect information is 0.95638 AUC value and 222 +- 566 mean ranking. After adding the side-effect information, I obtained the 0.9652 AUC value and 177+-513 mean ranking. Tests with only disease-drug networks without the side-effect information obtained a 0.9738 AUC value and 134 +-438 mean ranking. After incorporating the side-effect information, I obtained the 0.97411 AUC value and 132 +- 435 mean ranking. Therefore, the method achieves the best performance in drug repositioning when drug-disease relationships and side-effect information are considered.

Conclusion

I started our system design development with reproducing with our database which is used in the Wiki-Pi. First, by using ProphNet as our resource, I implement the converter to disease between OMIM and KEGG based on the disease name. I used string matching library to convert from OMIM to KEGG disease Codes. However, the AUC was decreased. The issue I faced was the less coverage for OMIM to KEGG disease codes. I only can obtain 321 pair of drug disease relation as a test set, even though the number of drugs were increased from 1337 to 6792. This huge difference of drug dataset and the test set might result in the lower AUC.

Also, I applied side effect similarity information into the original database. I implemented the algorithm for side effect similarity from the very first, since there was no reference codes. This work took a long time, but I finally finished the implementation. Finally, adding side effect similarity was also helpful to improve the performance. I obtained a 0.97411 AUC value and 132+-435 mean ranking.

Overall, through the drug repositioning wok, I learnt the domain knowledge and network based algorithms. Also, I have learnt so much about the state of art approaches in silico drug repositioning methods from the literature reviews.

References

1) Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321: 263–266.
2) Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Molecular systems biology 6: 343.
3) M. Gerstein, E. L. Sonnhammer, C. Chothia, *J Mol Biol.* 236, 1067 (1994).
4) Martınez, Vıctor, et al. "Network-based drug-disease relation prioritization using ProphNet.
5) Smith, Richard B. "Repositioned drugs: integrating intellectual property and regulatory strategies." Drug Discovery Today: Therapeutic Strategies 8.3 (2012): 131-137.
6) Loging, William, et al. "Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing." Drug Discovery Today: Therapeutic Strategies 8.3 (2012): 109-116.
7) Alaimo, Salvatore, et al. "Drug-Target interaction prediction through Domain-Tuned Network Based Inference." Bioinformatics (2013).
8) van Laarhoven, Twan, Sander B. Nabuurs, and Elena Marchiori. "Gaussian interaction profile kernels for predicting drug–target interaction." Bioinformatics 27.21 (2011): 3036-3043.
9) Wang, Yuhao, and Jianyang Zeng. "Predicting drug-target interactions using restricted Boltzmann machines." Bioinformatics 29.13 (2013): i126-i134.
10) Mei, Jian-Ping, et al. "Drug–target interaction prediction by learning from local information and neighbors." Bioinformatics 29.2 (2013): 238-245.
11) Takarabe, Masataka, et al. "Drug target prediction using adverse event report systems: a pharmacogenomic approach." Bioinformatics 28.18 (2012): i611-i618.
12) Mizutani, Sayaka, et al. "Relating drug–protein interaction network with drug side effects."
13) Bioinformatics 28.18 (2012): i522-i528.
14) Gottlieb, Assaf, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine." *Molecular systems biology* 7.1 (2011).