(Side-)Effect Prediction of Drugs

Mengyu Zhou, IIIS, Tsinghua

# Problem

## Feature Selection and Generation

## Multi-task Prediction

## Interpretive Pattern Discovery

# Data

## Drugbank

<http://www.drugbank.ca/>

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains 6729 drug entries including 1465 FDA-approved small molecule drugs, 132 FDA-approved biotech (protein/peptide) drugs, 86 nutraceuticals and 5076 experimental drugs. Additionally, 4254 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 150 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

## SIDER

<http://sideeffects.embl.de/>

SIDER contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. The available information include side effect frequency, drug and side effect classifications as well as links to further information, for example drug–target relations.

The current version has been released on October 17, 2012. This release uses the MedDRA dictionary (version 14.0) and provides access to preferred terms and lower-level terms. The number of drugs has increased from 888 to 996. There are 4192 side effects and 99423 drug-SE pairs. Compared to the release in March 2012, additional side effects have been retrieved by better processing of the labels. Side effects that are mentioned on the label as either potential or not occurring are removed.

## KEGG

<http://www.genome.jp/kegg/>

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

# Code

<http://github.com/zmy/EffectPredict>

## Extractors

Extractors are under **src/extractor**. They are the readers of original data. Now the extractors for DrugBank (local files), SIDER (local files) and KEGG (through the Internet) are well written. You can use the classes provided to get data easily.

## Experiment Code

Under **src/exp** there are some experimental code. Most of the code was written for extract a binary feature vector for drugs both exists in DrugBank and SIDER (800+). So there are mainly two parts: drug matching between DrugBank and SIDER based on names; and feature vector generation. You can start with Raw2Local.java.