Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Brief history of drug discovery and development

• Empirical – up until 1960’s
  – 14th–11th centuries BCE: herbal drugs, serendipitous discoveries
  – Late 1800’s: major pharmaceutical companies, mass production
  – 1920’s, 30’s: vitamins, vaccines
  – 1930-1960: major discoveries (insulin, penicillin, …)

• Rational – 1960’s to 1990’s
  – Designing molecules to target protein active sites – “lock and key”
  – Computational drug discovery
  – Biggest success HIV (RT, protease inhibitors)

• Big Experiment – 1990’s to 2000’s
  – High throughput screening
  – Microarray assays
  – Gene sequencing and human genome project

• Big Data – 2010’s onwards
  – Informatics-driven drug discovery
  – Everything is connected
Stages in the drug discovery process

Timescale in the drug discovery process

Bottleneck in drug discovery

Available at http://www.pharmafocusasia.com/strategy/drug_discovery_india_force_to_reckon.htm
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Big Data in the public domain

• There is now an incredibly rich resource of public information relating compounds, targets, genes, pathways, and diseases. Just for starters there is in the public domain information on*:
  – 48,777,362 compounds, 127,906,628 substances, 739,657 bioassays (PubChem)
  – 1552 FDA-approved small molecule drugs, 284 biotech drugs, 6009 experimental drugs (DrugBank)
  – 542,258 manually reviewed protein sequences, 51,616,950 un-reviewed protein sequences (Swiss-Prot/UniProtKB), 95,968 3D structures (PDB)
  – 22 million life science publications – 1 million new each year (PubMed)
  – 160,781 clinical studies with locations in all 50 states and in 185 countries (ClinicalTrials.gov)

• Even more important are the relationships between these entities. For example a chemical compound can be linked to a gene or a protein target in a multitude of ways:
  – Crystal structure of ligand/protein complex
  – Co-occurrence in a paper abstract
  – Computational experiment (docking, predictive model)
  – System association (e.g. involved in same pathways cellular processes)
  – Statistical relationship

* All databases were accessed on 02/08/2014
Why Data Mining is appealing

Buchan NS et al. Drug Discov Today. 2011 May;16(9-10):426-34.
**Why Drug Discovery and Development is appealing**

- Drug discovery is highly data driven and data are increasingly becoming public available
  - NIH has started ambitious extramural funding programs to support academic-based drug discovery programs recently
  - Pharms begin to make the trove of detailed raw data underlying its clinical trials systematically available to researchers
- Having ample data, bring challenging problems, demanding more knowledge
- Spans full data analytics cycles
  - Data collection, data cleansing, data semantics, data integration, data representation
  - Model inference, model selection, modal average, model interpretation
- We see many different data types
  - Vector, semi-structured, time-series, spatial-temporal, images, video, hypertext, literature
- Data analytics and data management challenges are from all aspects
  - Large volume, high dimensional, high noise, large amount of missing values, non iid data, structured input and output, unlabeled data
  - Multi-instance (label, class, task)
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
EHR data collection and analysis

Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Diagnosis data - ICD codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10
- Pros: Universally available
- Cons: medium recall and medium precision for characterizing patients

**Hypertensive disease (401 - 405)**

- (401) Essential hypertension
  - (401.0) Hypertension, malignant
  - (401.1) Hypertension, benign
  - (401.9) Hypertension, Unspecified
- (402) Hypertensive heart disease
- (403) Hypertensive renal disease
  - (403.0) Malignant hypertensive renal disease
  - (403.1) Benign hypertensive renal disease
- (404) Hypertensive heart and renal disease
- (405) Secondary hypertension
  - (405.0) Malignant secondary hypertension
    - (405.01) Hypertension, renovascular, malignant
  - (405.1) Benign secondary hypertension
    - (405.11) Hypertension, renovascular benign
Procedure data - CPT codes

- CPT stands for Current Procedural Terminology created by the American Medical Association
- CPT is used for billing purposes for clinical services
- Pros: High precision
- Cons: Low recall

Codes for surgery: 10021 - 69990

- (10021 - 10022) general
- (10040 - 19499) integumentary system
- (20000 - 29999) musculoskeletal system
- (30000 - 32999) respiratory system
- (33010 - 37799) cardiovascular system
- (38100 - 38999) hemic and lymphatic systems
- (39000 - 39599) mediastinum and diaphragm
- (40490 - 49999) digestive system
- (50010 - 53899) urinary system
- (54000 - 55899) male genital system
- (55920 - 55980) reproductive system and intersex
- (56405 - 58999) female genital system
- (59000 - 59899) maternity care and delivery
- (60000 - 60699) endocrine system
- (61000 - 64999) nervous system
- (65091 - 68899) eye and ocular adnexa
- (69000 - 69979) auditory system
Lab results

- The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®)

- Challenges for lab
  - Many lab systems still use local dictionaries to encode labs
  - Diverse numeric scales on different labs
    - Often need to map to normal, low or high ranges in order to be useful for analytics
  - Missing data
    - not all patients have all labs
    - The order of a lab test can be predictive, for example, BNP indicates high likelihood of heart failure

Hematology
ABG Analysis

Specimen: Arterial blood
Date and time specimen gathered: 07/21/2010 21:42pm

Blood Gases:

<table>
<thead>
<tr>
<th>Acid/Base</th>
<th>Results</th>
<th>Reference Range</th>
<th>Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.27</td>
<td>7.35-7.45</td>
<td>(L)</td>
</tr>
<tr>
<td>pCO2</td>
<td>48 mmHg</td>
<td>35-45 mmHg</td>
<td></td>
</tr>
<tr>
<td>pO2</td>
<td>92 mmHg</td>
<td>80-100 mmHg</td>
<td>(H)</td>
</tr>
<tr>
<td>HCO3</td>
<td>25 mEq/L</td>
<td>24-26 mEq/L</td>
<td></td>
</tr>
<tr>
<td>O2 sat</td>
<td>97%</td>
<td>95-100%</td>
<td></td>
</tr>
</tbody>
</table>
Medication

- Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug
  - Not used universally by EHR systems
  - Too specific, drugs with the same ingredients but different brands have different NDC

- RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine

- Medication data can vary in EHR systems
  - can be in both structured or unstructured forms

- Availability and completeness of medication data vary
  - Inpatient medication data are complete, but outpatient medication data are not
  - Medication usually only store prescriptions but we are not sure whether patients actually filled those prescriptions
Clinical notes

• Clinical notes contain rich and diverse source of information
• Challenges for handling clinical notes
  – Ungrammatical, short phrases
  – Abbreviations
  – Misspellings
  – Semi-structured information
    • Copy-paste from other structure source
      – Lab results, vital signs
    • Structured template:
      – SOAP notes: Subjective, Objective, Assessment, Plan

Client arrived to discuss previously established goal:

Reduce psychological energy and return to premorbid levels of activity, judgment, mood, and goal-directed behavior.

Elmira reported that her speech rate increases as she feels stressed.
## Strengths and weakness of data classes within EHRs

<table>
<thead>
<tr>
<th></th>
<th>ICD codes</th>
<th>CPT codes</th>
<th>Laboratory Data</th>
<th>Medication records</th>
<th>Clinical Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Availability in EHR</strong></td>
<td><strong>Near-universal</strong></td>
<td><strong>Near-universal</strong></td>
<td><strong>Near-universal</strong></td>
<td><strong>Variable</strong></td>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td><strong>Medium</strong></td>
<td><strong>Poor</strong></td>
<td><strong>Medium</strong></td>
<td><strong>Inpatient: High</strong></td>
<td><strong>Outpatient: Variable</strong></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td><strong>Medium</strong></td>
<td><strong>High</strong></td>
<td><strong>High</strong></td>
<td><strong>Inpatient: High</strong></td>
<td><strong>Outpatient: Variable</strong></td>
</tr>
<tr>
<td><strong>Fragmentation effect</strong></td>
<td><strong>Medium</strong></td>
<td><strong>High</strong></td>
<td><strong>Medium-High</strong></td>
<td><strong>Medium</strong></td>
<td><strong>Low-Medium</strong></td>
</tr>
<tr>
<td><strong>Query method</strong></td>
<td><strong>Structured</strong></td>
<td><strong>Structured</strong></td>
<td><strong>Mostly structured</strong></td>
<td><strong>Structured, text</strong></td>
<td><strong>queries, and NLP</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>NLP, text queries, and</strong></td>
<td><strong>rarely structured</strong></td>
</tr>
<tr>
<td><strong>Strengths</strong></td>
<td>-Easy to query</td>
<td>-Easy to query</td>
<td>-Value depends on test</td>
<td>Can have high validity</td>
<td>Best record of what providers thought</td>
</tr>
<tr>
<td></td>
<td>-Serves as a good first pass of disease status</td>
<td>-High precision</td>
<td>-High data validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Weaknesses</strong></td>
<td>-Disease codes often used for screening when disease not actually present</td>
<td>-Most susceptible to missing data errors (e.g., performed at another hospital)</td>
<td>-May need to aggregate different variations of the same data elements</td>
<td>-Often need to interface inpatient and outpatient records</td>
<td>-Difficult to process automatically</td>
</tr>
<tr>
<td></td>
<td>-Accuracy hindered by billing realities and clinic workflow</td>
<td>-Procedure receipt influenced by patient and payer factors external to disease process</td>
<td>-Normal ranges and units may change over time</td>
<td>-Medication records from outside providers not present</td>
<td>-Interpretation accuracy depends on assessment method</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-Medications prescribed not necessary taken</td>
<td>-May suffer from significant “cut and paste”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-Not universally available in EHRs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-May be self-contradictory</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>Essential first element for electronic phenotyping</td>
<td>Helpful addition if relevant</td>
<td>Helpful addition if relevant</td>
<td>Useful for confirmation and a marker of severity</td>
<td>Useful for confirming common diagnoses or for finding rare ones</td>
</tr>
</tbody>
</table>

EHR data description and patient vector represent:

- **Diagnosis**
  - ICD9
  - CCS hierarchy
  - HCC hierarchy
  - co-occurring HCC

- **Procedure**
  - CPT
  - CPT CCS hierarchy
  - RVU as value

- **Pharmacy**
  - NDC
  - Ingredient
  - Days of Supplies

- **Lab**
  - Lab results
  - Break down by age and sex groups

- **Demography**
  - Age
  - Gender

The extracted features are represented as:

\[ x_1, x_2, \ldots, x_N \]

Patient Feature Vector
Moving towards to Personalized Medicine

- Personalized Medicine: the right patient with the right drug at the right dose at the right time.
- Safer, more effective drugs: end of one-size fits-all drugs. Target discoveries enable development of drugs that will be safer and effective for specific populations.
- Faster time to market: using genomic and real-world data to find disease targets. Speedier clinical trials based on high responder population.
- Cost-effective healthcare: reduced costs, due to avoidance of futile treatments and improved clinical outcomes. Better treatment adherence = increased profitability.
Intuition of Personalized Medicine methods

- Combine patient similarity with drug similarity analysis.
- Leverages large amount of real-world data available for "mature" drugs to derive information relevant for a new drug.
Network based approach combining drug similarity with patient similarity

- Construct a heterogeneous patient-drug graph $A$:

$$A = \begin{bmatrix} S_p & R \\ R^T & S_d \end{bmatrix}$$

- Spread the information representing the effectiveness of different drugs for different patients (vector $Y$) by a label propagation algorithm:

$$F = (1-\mu)(I-\mu W)^{-1}Y$$

Zhang P et al. Summit on Translational Bioinformatics (TBI), 2014.
Application: personalized treatments for hypercholesterolemia

Data: 1219 distinct patients and 4 statin cholesterol-lowering drugs from a real-world EHR

<table>
<thead>
<tr>
<th>Drug</th>
<th>Patient #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atorvastatin</td>
<td>97</td>
</tr>
<tr>
<td>Lovastatin</td>
<td>221</td>
</tr>
<tr>
<td>Pravastatin</td>
<td>24</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>877</td>
</tr>
</tbody>
</table>

90-day patient diagnosis condition assessment period

First drug day during the 60-day window

60-day window use a single drug

Consecutive 2 in-control check points
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Examples of drug repositioning

*New uses for old drugs*

<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>
Meet the unmet medical needs efficiently

Dependent and Independent Variables in Drug Repositioning

- **Indication (Y)**
- **Target ID**
- **Preclinical**
- **Clinical**
- **Launch**

**Pharmaco-information (X)**

- **Launch**
  - (Pharmacological effect on human)
- **Clinical Trial**
  - (Pharmacological effect on human)
- **Animal Model**
  - (Pharmacological effects on animal)
- **Compound Profiling**
  - (Gene Expression, Phenotypic Screen, Structure)
- **Target**
  - (On-target, Off-target)

\[ Y \text{ (indication)} = f \left( X_1, X_2, \ldots, X_n \right) \]
Identify the off-targets via Chemical-Protein Interactome (CPI)

• Introduction of the CPI

• Case study
  – Clozapine induced agranulocytosis (CIA)
    • Although agranulocytosis is a side effect, the methodology is applicable to the identification of the therapeutic effect
Chemical-protein interactions

The DOCK

- A program used to simulate the chemical-protein interactions and to measure the interaction strength

- Provide the theoretical binding conformation of the drug’s binding to protein

- A lower docking score means a higher binding strength

\[ E_{\text{inter}} = \sum_{i=1}^{k_g} \sum_{j=1}^{r_{ij}} \left( \frac{A_{ij}}{r_{ij}^n} - \frac{B_{ij}}{r_{ij}^m} + 3.820 \cdot 10^{-3} \cdot \frac{g_{ij}}{D_{ij}} \right), \]

van der Waals and electrostatic interaction

http://dock.compbio.ucsf.edu/
Binding conformation in Chemical-Protein Interactome (CPI)

Direct binding model of sulfonamides - MHC I (Cw*4) interactions

Binding strength in CPI

Two Directional Z-transformation (2DIZ) of Docking Scores $X_{ij}$

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{SD_{X_j}}$$

$$Z'_{ij} = \frac{Z_{ij} - \bar{Z}_i}{SD_{Z_i}}$$

Linear Model of the Docking Scores

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

$$b = \frac{\sum_{q=1}^{n} \sum_{k=1}^{m} (\alpha\beta)_{kq}}{mn}$$

$$Z'_{ij} = \frac{-b\sqrt{n-1}}{\sqrt{(n-1)b^2 + [(z\beta)_i - b]^2}} \quad (i \neq j), \quad \text{when } (m \rightarrow +\infty, n \rightarrow +\infty)$$

$$Z'_{ij} = \left[(z\beta)_{ij} - b\right] \sqrt{\frac{(n-1)}{(n-1)b^2 + [(z\beta)_{ij} - b]^2}} \quad (i = j),$$

when $(m \rightarrow +\infty, n \rightarrow +\infty)$,

Yang, L. et al. PLoS ONE. 2010
Improve the performance of the docking scores via using 2DIZ

Benchmark structural model set:
100 pockets with their embedded ligands

High variability in ligand structures

<table>
<thead>
<tr>
<th>Test Result Variable(s)</th>
<th>AUC</th>
<th>Std. Errora</th>
<th>Asymptotic Sig.b</th>
<th>Asymptotic 95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docking Score</td>
<td>.623</td>
<td>.033</td>
<td>.000</td>
<td>Lower Bound: .558, Upper Bound: .687</td>
</tr>
<tr>
<td>Z-score</td>
<td>.759</td>
<td>.028</td>
<td>.000</td>
<td>Lower Bound: .703, Upper Bound: .815</td>
</tr>
<tr>
<td>Z'-score</td>
<td>.823</td>
<td>.021</td>
<td>.000</td>
<td>Lower Bound: .781, Upper Bound: .866</td>
</tr>
</tbody>
</table>
Identify the True Chemical-Protein Interactions

<table>
<thead>
<tr>
<th>Ligands</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Rank</td>
<td>Low Rank</td>
</tr>
</tbody>
</table>

- **Known Ligand-target Pair**
- **Negative**
False Positive - Tolerant MCC (FPT-MCC)

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

\[
|MCC| = \sqrt{\frac{\chi^2}{n}}
\]

\[
(FPT-MCC) = \frac{TP' \times TN - FP' \times FN}{\sqrt{(TP' + FP')(TP' + FN)(TN + FP')(TN + FN)}}
\]

\[
TP' = TP + \alpha FP
\]

\[
FP' = (1-\alpha)FP
\]
<table>
<thead>
<tr>
<th>Class</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>100</td>
<td>10,000</td>
</tr>
<tr>
<td>Mean</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>St. Dev</td>
<td>1.5</td>
<td>1</td>
</tr>
</tbody>
</table>

AUC = 0.675
Sn = 0.73
Sp = 0.51
Real Sp > 0.51

(A) (B) (D) (C)

FPT-MCC
MCC

Z’-score

Fraction

Z’-score

FPR

TPR
Case Study:
Identify Off-targets for Clozapine induced agranulocytosis (CIA)

Clozapine (CLZ)

Olanzapine (OLZ)

Causing fatal agranulocytosis -- A deficiency of granulocytes in the blood

The difference of the agranulocytosis report rate between clozapine and olanzapine in the FDA AERS

<table>
<thead>
<tr>
<th></th>
<th>Clozapine</th>
<th>Olanzapine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agranulocytosis Reports</td>
<td>185</td>
<td>16</td>
</tr>
<tr>
<td>Total Reports</td>
<td>16813</td>
<td>11304</td>
</tr>
<tr>
<td>Ratio of Agranulocytosis Report (%)</td>
<td>1.1</td>
<td>0.14</td>
</tr>
<tr>
<td>$P_{CLZ-OLZ}^*$</td>
<td>1.1</td>
<td>8.2E-21</td>
</tr>
</tbody>
</table>
Identifying off-targets for CIA

Resource Specifications for Docking

- Blue Meadow cluster
  - Located at Ashburn contractor operated data center
  - IBM iDataPlex dx360 M2 Server machines & Sun Grid Engine, PBS
  - 252 nodes x 8 cores = 2016 cores
  - 6TB RAM, or 24 GB per node
  - Memory distributed between nodes & shared within nodes
ANOVA of the chemical-protein interactive effect before and after 2DIZ

### Before 2DIZ

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>F</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>409</td>
<td>2333527</td>
<td>111.22</td>
<td>&lt;2.2e-16</td>
</tr>
<tr>
<td>Chemical</td>
<td>254</td>
<td>10330585</td>
<td>793.27</td>
<td>&lt;2.2e-16</td>
</tr>
<tr>
<td>Interactive</td>
<td>95344</td>
<td>4888387</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### After 2DIZ

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>F</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>409</td>
<td>0</td>
<td>1.37E-19</td>
<td>1</td>
</tr>
<tr>
<td>Chemical</td>
<td>254</td>
<td>1052</td>
<td>4.1776</td>
<td>&lt;2.2e-16</td>
</tr>
<tr>
<td>Interactive</td>
<td>95344</td>
<td>94546</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Protein Effect has been excluded. May use drug to fish proteins.
- Normalize the chemical effect first.
Candidate off-targets from Binomial Antithesis CPI

<table>
<thead>
<tr>
<th>PDB ID*</th>
<th>Target Name</th>
<th>Gene Name</th>
<th>$Z^<em>$ (CLZ)</em></th>
<th>$Z^*$ (OLZ)</th>
<th>A-score</th>
<th>p value for CPI</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CBS</td>
<td>Cellular retinoic acid-binding protein 2</td>
<td>CRABP2</td>
<td>-0.922</td>
<td>1.653</td>
<td>-2.575</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>1D1T</td>
<td>Alcohol dehydrogenase class 4 mu/sigma chain</td>
<td>ADH7</td>
<td>-1.191</td>
<td>1.525</td>
<td>-2.716</td>
<td>0.000</td>
<td>OR</td>
</tr>
<tr>
<td>1HL_1</td>
<td>Aldo-keto reductase family 1 member C2</td>
<td>AKR1C2</td>
<td>-0.781</td>
<td>2.545</td>
<td>-3.326</td>
<td>0.000</td>
<td>OR</td>
</tr>
<tr>
<td>1HL_2</td>
<td>Aldo-keto reductase family 1 member C2</td>
<td>AKR1C2</td>
<td>-1.605</td>
<td>1.023</td>
<td>-2.628</td>
<td>0.000</td>
<td>OR</td>
</tr>
<tr>
<td>1O1Z</td>
<td>Alpha-tocopherol transfer protein</td>
<td>TTPA</td>
<td>-1.269</td>
<td>1.171</td>
<td>-2.440</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>2E8A</td>
<td>Heat shock 70 kDa protein 1</td>
<td>HSPA1A/HSPA1B</td>
<td>-1.381</td>
<td>0.150</td>
<td>-1.531</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>1D2V</td>
<td>Myeloperoxidase</td>
<td>MPO</td>
<td>-2.753</td>
<td>-0.646</td>
<td>-2.107</td>
<td>0.005</td>
<td>OR</td>
</tr>
<tr>
<td>1DB1</td>
<td>Vitamin D3 receptor</td>
<td>VDR</td>
<td>-0.660</td>
<td>0.748</td>
<td>-1.409</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>1MRQ_2</td>
<td>Aldo-keto reductase family 1 member C1</td>
<td>AKR1C1</td>
<td>-2.034</td>
<td>0.123</td>
<td>-2.158</td>
<td>0.016</td>
<td>OR</td>
</tr>
<tr>
<td>1MRQ_1</td>
<td>Aldo-keto reductase family 1 member C1</td>
<td>AKR1C1</td>
<td>-1.036</td>
<td>0.601</td>
<td>-1.637</td>
<td>0.021</td>
<td>OR</td>
</tr>
<tr>
<td>1DHT</td>
<td>Estradiol 17-beta-dehydrogenase 1</td>
<td>HSD17B1</td>
<td>-1.822</td>
<td>0.158</td>
<td>-1.980</td>
<td>0.021</td>
<td>OR</td>
</tr>
<tr>
<td>1MUO</td>
<td>Serine/threonine-protein kinase 6</td>
<td>AURKA</td>
<td>-1.136</td>
<td>0.529</td>
<td>-1.665</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>1VJ5</td>
<td>Epoxide hydrolase 2</td>
<td>EPHX2</td>
<td>-1.088</td>
<td>0.228</td>
<td>-1.315</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>4GTU</td>
<td>Glutathione S-transferase Mu 4</td>
<td>GSTM4</td>
<td>-0.749</td>
<td>1.060</td>
<td>-1.809</td>
<td>0.036</td>
<td>GT</td>
</tr>
<tr>
<td>1HDR</td>
<td>Dihydropyrimidine reductase</td>
<td>QDPR</td>
<td>-1.469</td>
<td>0.561</td>
<td>-2.030</td>
<td>0.038</td>
<td>OR</td>
</tr>
<tr>
<td>1YB5</td>
<td>Quinone oxidoreductase</td>
<td>CRYZ</td>
<td>-1.212</td>
<td>0.284</td>
<td>-1.496</td>
<td>0.039</td>
<td>OR</td>
</tr>
<tr>
<td>1CM8</td>
<td>Mitogen-activated protein kinase 12</td>
<td>MAPK12</td>
<td>-1.202</td>
<td>0.301</td>
<td>-1.503</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td>1XF0_2</td>
<td>Aldo-keto reductase family 1 member C3</td>
<td>AKR1C3</td>
<td>-0.865</td>
<td>0.441</td>
<td>-1.306</td>
<td>0.041</td>
<td>OR</td>
</tr>
<tr>
<td>1HMR</td>
<td>Fatty acid-binding protein, heart</td>
<td>FABP3</td>
<td>-0.826</td>
<td>0.270</td>
<td>-1.095</td>
<td>0.046</td>
<td></td>
</tr>
</tbody>
</table>

*An entry name that ends with a number represents the pocket number of its PDB structure.

*The smaller $Z^*$-score represents a higher theoretical interaction strength.

In the “Role” column, OR and GT indicate oxidoreductases and glutathione metabolism related proteins, respectively.
Hsp70 protein is the candidate off-target of CLZ not OLZ

<table>
<thead>
<tr>
<th>PDB ID#</th>
<th>Target Name</th>
<th>Gene Name</th>
<th>Z’ (CLZ)*</th>
<th>Z’ (OLZ)</th>
<th>A-score</th>
<th>p value for CPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>2E8A</td>
<td>Heat shock 70 kDa protein 1</td>
<td>HSPA1A/HSPA1B</td>
<td>−1.381</td>
<td>0.150</td>
<td>−1.531</td>
<td>0.001</td>
</tr>
</tbody>
</table>

CLZ: Clozapine

OLZ: Olanzapine
Candidate off-targets prioritized from Multiple Antithesis CPI

<table>
<thead>
<tr>
<th>PDB ID*</th>
<th>Target Name</th>
<th>Gene Name</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>RR</th>
<th>p value</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1110</td>
<td>L-lactate dehydrogenase A chain</td>
<td>LDHA</td>
<td>21</td>
<td>1</td>
<td>18</td>
<td>14</td>
<td>1.697</td>
<td>0.002</td>
<td>OR</td>
</tr>
<tr>
<td>2HG5_2</td>
<td>Glutathione synthetase</td>
<td>GSS</td>
<td>24</td>
<td>2</td>
<td>15</td>
<td>13</td>
<td>1.723</td>
<td>0.002</td>
<td>GT</td>
</tr>
<tr>
<td>2HRB</td>
<td>Carbonyl reductase NAD(P)H 3</td>
<td>CBR3</td>
<td>17</td>
<td>0</td>
<td>22</td>
<td>15</td>
<td>1.662</td>
<td>0.002</td>
<td>OR</td>
</tr>
<tr>
<td>1KBQ</td>
<td>NAD(P)H dehydrogenase quinone 1</td>
<td>NQO1</td>
<td>16</td>
<td>0</td>
<td>23</td>
<td>15</td>
<td>1.652</td>
<td>0.002</td>
<td>OR</td>
</tr>
<tr>
<td>1EEM</td>
<td>Glutathione S-transferase omega-1</td>
<td>GSTO1</td>
<td>19</td>
<td>1</td>
<td>20</td>
<td>14</td>
<td>1.615</td>
<td>0.004</td>
<td>GT</td>
</tr>
<tr>
<td>1SG0_2</td>
<td>Ribosylhydronicotinamide dehydrogenase quinone</td>
<td>NQO2</td>
<td>14</td>
<td>0</td>
<td>22</td>
<td>15</td>
<td>1.662</td>
<td>0.005</td>
<td>OR</td>
</tr>
<tr>
<td>1G0X</td>
<td>Leukocyte immunoglobulin-like receptor subfamily B member 1</td>
<td>LILRB1</td>
<td>15</td>
<td>0</td>
<td>24</td>
<td>15</td>
<td>1.625</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>2AHE</td>
<td>Chloride intracellular channel protein 4</td>
<td>CLIC4</td>
<td>14</td>
<td>0</td>
<td>24</td>
<td>15</td>
<td>1.625</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>1DIA</td>
<td>Formyltetrahydrofolate synthetase</td>
<td>MTHFD1</td>
<td>14</td>
<td>0</td>
<td>25</td>
<td>15</td>
<td>1.600</td>
<td>0.006</td>
<td>OR</td>
</tr>
<tr>
<td>11GS</td>
<td>Glutathione S-transferase P</td>
<td>GSTP1</td>
<td>18</td>
<td>1</td>
<td>21</td>
<td>14</td>
<td>1.579</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>1FIE</td>
<td>Coagulation factor XIII A chain</td>
<td>F13A1</td>
<td>12</td>
<td>0</td>
<td>23</td>
<td>15</td>
<td>1.652</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>1Q4O</td>
<td>Serine/threonine-protein kinase PLK1</td>
<td>PLK1</td>
<td>13</td>
<td>0</td>
<td>25</td>
<td>15</td>
<td>1.600</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>1LJR</td>
<td>Glutathione S-transferase theta-2</td>
<td>GSTT2B</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>14</td>
<td>2.167</td>
<td>0.011</td>
<td>GT</td>
</tr>
<tr>
<td>1FPR</td>
<td>Tyrosine-protein phosphatase non-receptor type 6</td>
<td>PTPN6</td>
<td>13</td>
<td>0</td>
<td>26</td>
<td>15</td>
<td>1.577</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>1HSO</td>
<td>Alcohol dehydrogenase 1A</td>
<td>ADH1A</td>
<td>13</td>
<td>0</td>
<td>26</td>
<td>15</td>
<td>1.577</td>
<td>0.011</td>
<td>OR</td>
</tr>
<tr>
<td>1IH1_1</td>
<td>Aldo-keto reductase family 1 member C2</td>
<td>AKR1C2</td>
<td>20</td>
<td>2</td>
<td>19</td>
<td>13</td>
<td>1.531</td>
<td>0.014</td>
<td>OR</td>
</tr>
<tr>
<td>1SG0_1</td>
<td>Ribosylhydronicotinamide dehydrogenase quinone</td>
<td>NQO2</td>
<td>16</td>
<td>1</td>
<td>22</td>
<td>14</td>
<td>1.540</td>
<td>0.020</td>
<td>OR</td>
</tr>
<tr>
<td>1IH1_2</td>
<td>Aldo-keto reductase family 1 member C2</td>
<td>AKR1C2</td>
<td>16</td>
<td>1</td>
<td>23</td>
<td>14</td>
<td>1.514</td>
<td>0.021</td>
<td>OR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Agranulocytosis +</th>
<th>Agranulocytosis-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not Binding</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>
**Rationale of Using Pharmacological Effects in Drug Repositioning**

**Side-effects (SE) and therapeutic effects**

- **Side-effects (SE)** and therapeutic effects are clinical phenotypic effects of drug treatment
  - They may associate with each other via underlying mechanism

**Clinical phenotypic information comes from patients, not animals**

- Mimics a human phenotypic ‘assay’
- May have less translational issue

**Quantitative Rational**

\[
\max(P(D_i \mid se_1, se_2, \ldots, se_m)), \ i \in ([D]) \\
\text{posterior} \\
P(D_i \mid se_1, se_2, \ldots, se_m) = \frac{P(se_1, se_2, \ldots, se_m \mid D_i)P(D_i)}{P(se_1, se_2, \ldots, se_m)}
\]

\[
P(se_1, se_2, \ldots, se_m \mid D_i) = \prod_{j=1}^{m} P(se_j \mid D_i)
\]

• Identification of the disease-side effect associations
Retrieving side-effect/disease information from drug label and PharmGKB

Side Effect

Skin:
Allergic skin reactions, e.g., pruritus, erythema, urticaria, vasculitis, and morbilliform or maculopapular eruptions, occur in less than 1% of treated patients. Such mild reactions may develop into serious reactions sometimes progressing to shock. These may be transient and may disappear despite continued use of glimepiride if skin reactions persist, the drug should be discontinued. Although there have been no reports for glimepiride, porphyria cutanea tarda

Identification of the disease-side effect associations

584 side effects; 145 diseases; 3175 informative drug-SE associations

Yang, L. Agarwal, P. PLoS ONE, 2011
## Examples of disease-side effect associations

<table>
<thead>
<tr>
<th>Disease Class</th>
<th>Disease</th>
<th>Side Effect</th>
<th>MCC</th>
<th>sn</th>
<th>sp</th>
<th>p value</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulation</td>
<td>Stroke</td>
<td>Positive ANA</td>
<td>0.46</td>
<td>0.47</td>
<td>0.98</td>
<td>1.8E-15</td>
<td>statins, ramipril</td>
</tr>
<tr>
<td>Immune System</td>
<td>Transplant rejection</td>
<td>Cytomegalovirus infection</td>
<td>0.75</td>
<td>0.75</td>
<td>0.99</td>
<td>3.5E-06</td>
<td>methotrexate</td>
</tr>
<tr>
<td>Metabolite disease</td>
<td>Diabetes Mellitus</td>
<td>Porphyria</td>
<td>0.44</td>
<td>0.50</td>
<td>0.98</td>
<td>8.8E-06</td>
<td>valproic acid, pyrazinamide, naproxen, estradiol</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>Depressive Disorder</td>
<td>Delusions</td>
<td>0.46</td>
<td>1.00</td>
<td>0.91</td>
<td>1.1E-08</td>
<td>cabergoline, memantine, pergolide</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>Depressive Disorder</td>
<td>Hyperacusis</td>
<td>0.55</td>
<td>0.88</td>
<td>0.96</td>
<td>9.0E-09</td>
<td>phenytoin, modafinil</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>Neoplasms</td>
<td>Constitutional symptoms</td>
<td>0.50</td>
<td>0.56</td>
<td>0.94</td>
<td>2.6E-18</td>
<td>nevirapine</td>
</tr>
</tbody>
</table>

Diagram:

```
Transplant rejection —> Methotrexate

CMV infections
```
**Stroke - positive Antinuclear Antibodies (ANA)**

<table>
<thead>
<tr>
<th>Disease Class</th>
<th>Disease</th>
<th>Side Effect</th>
<th>MCC</th>
<th>sn</th>
<th>sp</th>
<th>p value</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulation</td>
<td>Stroke</td>
<td>Positive ANA</td>
<td>0.46</td>
<td>0.47</td>
<td>0.98</td>
<td>1.8E-15</td>
<td>statins, ramipril</td>
</tr>
<tr>
<td>System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The SE *positive ANA* is shared by drugs treating Stroke
  - mainly ticlopidine and several angiotensin-converting enzyme (ACE) inhibitors
- Stroke is associated with severe immune suppression
- Drugs that are associated with increasing immune response in terms of *positive ANA* may help stroke patients
- Ramipril lists *positive ANA* as a SE
  - showed a 32% risk reduction for stroke


Diabetes - *porphyria*

- A metabolic disease characterized by errors in the biosynthetic pathway of HEME

- A partial defect in the activity of the porphobilinogen deaminase during the HEME synthesis

- Carbohydrates modulate the HEME synthesis

- *Porphyria* is hereditary – this ‘biomarker’ only works on these knock-out people

The balancing between SE and disease

- Diabetic drug worsen *porphyria*
- *Porphyria* resolved after developing diabetes
  - 328 patients with *porphyria*, the 16 patients that developed diabetes all had their *porphyria* symptoms resolved
  - 16 “knock-out” people mimic a phenotypic screening for diabetic drug

Diabetes - *porphyria*

- Drugs list *porphyria* as a SE but are not indicated for diabetes could be tested for treating diabetes
  - Valproic acid is an anticonvulsant and a recent study found it effective in lowering blood glucose levels in mice
  - In mice, pain killer naproxen is used as a tool to delay or prevent the development of type II diabetes from a pre-diabetic condition

• Drug Repositioning based on Side Effects (DRoSEf) for marketed drugs
Repositioning using single SE feature

<table>
<thead>
<tr>
<th></th>
<th>Case: Parkinson +</th>
<th>Ctrl: Parkinson -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priapism +</td>
<td>10 (TP)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(FP, drugs listing priapism not used to treat PD)</td>
</tr>
<tr>
<td>Priapism -</td>
<td>4 (FN, drugs treating PD not inducing priapism)</td>
<td>271 (TN)</td>
</tr>
<tr>
<td>Sensitivity = 10 / (10 + 4) = 0.71</td>
<td>Specificity = 271 / (271 + 18) = 0.94</td>
<td></td>
</tr>
</tbody>
</table>

- 27% (16346) of the new drugs-disease association suggested by DRoSEf have at least one entry in PubMed
- 44194 repositioning opportunities for marketed drugs

D1, D2, ..., D18
AUCs of 10-fold cross validations across 145 diseases using multiple SE features

<table>
<thead>
<tr>
<th>Disease</th>
<th>AUC</th>
<th>Disease</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amyotrophic Lateral Sclerosis</td>
<td>1</td>
<td>Influenza, Human</td>
<td>0.997</td>
</tr>
<tr>
<td>Anemia</td>
<td>1</td>
<td>Leukemia, Lymphocytic, Chronic, B-Cell</td>
<td>1</td>
</tr>
<tr>
<td>Arthritis</td>
<td>1</td>
<td>Liver Neoplasms</td>
<td>1</td>
</tr>
<tr>
<td>Asthma</td>
<td>0.959</td>
<td>Migraine without Aura</td>
<td>1</td>
</tr>
<tr>
<td>Cough</td>
<td>0.998</td>
<td>Myopathy, Central Core</td>
<td>1</td>
</tr>
<tr>
<td>Dementia</td>
<td>1</td>
<td>Non-small cell lung cancer</td>
<td>0.986</td>
</tr>
<tr>
<td>Diabetic Nephropathies</td>
<td>1</td>
<td>normal tension glaucoma</td>
<td>1</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>0.982</td>
<td>Osteonecrosis</td>
<td>0.993</td>
</tr>
<tr>
<td>Esophageal Neoplasms</td>
<td>0.983</td>
<td>Osteoporosis, Postmenopausal</td>
<td>1</td>
</tr>
<tr>
<td>estrogen-dependent carcinogenesis</td>
<td>1</td>
<td>Pain</td>
<td>0.983</td>
</tr>
<tr>
<td>Gastroesophageal Reflux</td>
<td>0.997</td>
<td>Parkinson Disease</td>
<td>0.959</td>
</tr>
<tr>
<td>Glioblastoma</td>
<td>1</td>
<td>Peripheral Nervous System Diseases</td>
<td>0.957</td>
</tr>
<tr>
<td>Glomerulosclerosis</td>
<td>0.997</td>
<td>Psoriasis</td>
<td>0.962</td>
</tr>
<tr>
<td>Heart Diseases</td>
<td>1</td>
<td>Rectal Neoplasms</td>
<td>0.983</td>
</tr>
<tr>
<td>Hyperlipidemias</td>
<td>0.981</td>
<td>Rheumatic Diseases</td>
<td>0.994</td>
</tr>
</tbody>
</table>

92% of the AUCs were above 0.8
• DRoSEf for clinical molecules
Quantitative Structure Activity Relationship (QSAR) Modeling

• Drug-like properties
  – Octanol-water partition coefficient (logP)
  – Hydrogen bond donors
  – Hydrogen bond acceptors
  – Molecular Mass

• Structural Signature
DRoSEf for clinical molecules

- 4,200 clinical molecules that are indicated for at least one of 101 diseases

\[
DS_i = [d_{s_{i1}}, d_{s_{i2}}, ..., d_{s_{ij}}], \quad j \in [1, 566], \quad i \in [1, 101]
\]

\[
SM_k = [s_{m_{1k}}, s_{m_{2k}}, ..., s_{m_{jk}}], \quad j \in [1, 566], \quad k \in [1, 4200]
\]

\[
d_{s_{ij}} \in \{b_{ij}, mcc_{ij}, mcc_{ij}^4, s_{n_{ij}}, s_{n_{ij}}^4, s_{p_{ij}}, s_{p_{ij}}^4\}
\]

\[
\Theta_{ik} = \sum_{j=1}^{566} d_{s_{ij}} s_{m_{jk}} \rightarrow \Theta_{i2} > \Theta_{i1}
\]
**Prediction results for clinical molecules**

<table>
<thead>
<tr>
<th>Disease category</th>
<th>Disease</th>
<th># of drugs</th>
<th># of SE features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neuropsychiatric</strong></td>
<td>Depression</td>
<td>72</td>
<td>87</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Depressive Disorder</td>
<td>42</td>
<td>204</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Schizophrenia</td>
<td>77</td>
<td>55</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Depressive Disorder, Major</td>
<td>48</td>
<td>170</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Anxiety Disorders</td>
<td>144</td>
<td>186</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Neoplasms</strong></td>
<td>Stomach Neoplasms</td>
<td>49</td>
<td>4</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Carcinoma, Non-Small-Cell Lung</td>
<td>73</td>
<td>10</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Lung Neoplasms</td>
<td>59</td>
<td>30</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Neoplasms</td>
<td>347</td>
<td>42</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Lymphoma</td>
<td>28</td>
<td>4</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Leukemia</td>
<td>30</td>
<td>20</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Head and Neck Neoplasms</td>
<td>33</td>
<td>7</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td>Hypertension</td>
<td>203</td>
<td>12</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Diabetes Mellitus, Type 2</td>
<td>112</td>
<td>8</td>
<td>0.71</td>
</tr>
</tbody>
</table>

4200 Molecules * 101 Disease Endpoints

\[ ds_{ij} \in \{b_{ij}, mcc_{ij}, mcc_{ij}^+, sn_{ij}, sn_{ij}^+, sp_{ij}, sp_{ij}^+\} \]
Case Study: Predict drugs’ repositioning potential for hypertension
DRoSEf vs. QSAR alone

AUC comparison of different methods
Discussion

• Frequency information of the SE is not considered

• High frequency may not be informative
  – diarrhea, dizziness, Vomiting, Nausea

• Rare SE may be informative
  – Although some side effects are rare, but they can also
    • shed lights on the association between side effects and new indication
    • drugs are screened on the models using ‘knock-out’ animals
  – Just like rare mutations identified from the GWAS of common diseases could
    • shed lights on the pathogenesis of common diseases
Take-home message

• Drug indication can be suggested based on clinical side-effects
  – Mimicking a phenotypic efficacy screen on some “knock-out” human

• DRoSEf may also suggest the neglected pathogenesis of disease
  – For example, studying *porphyria* may help discover potential new mode of action for diabetes therapy
The Connectivity Map concept

(a) Drug-centred approach

- CMap drug
- Drug of interest
- Microarrays
  gene expression
- Similar
- ✓ Discovery of new actions
  ✓ Mechanisms of action

(b) Disease-centred approach

- CMap drug
- Human disease
  (tissue, cells)
- Microarrays
  gene expression
- Opposite
- ✓ Drug repositioning
  ✓ Suggest new treatments

What we are interested in

• Identify Associations among diseases (Y) and:
  – Chemical-protein interactome profile of drug
  – Gene expression profile after drug treatment
  – Side Effect of drug [e.g., Parkinson Disease and Priapism]

• Prediction
  – How to use the Pharmaco-information \((X_1, X_2, \ldots, X_n)\) to predict indication of it
  – Prediction results should be understood by biologists
Drug-disease Associations from various resources
Challenges

• Many factors need to be considered for practical use of drug repositioning
  – the unmet medical need for the disease
  – the CNS penetration of the molecule
  – therapeutic effect is significant compared with the active comparators
  – the previous therapeutic effect could now become a side effect
  – ...

• These issues may be controlled via choosing a suitable formulation, dose, and the sub population
Outline

• Introduction of Drug Discovery and Development
• Motivation of Data Mining
• Data Sources for Data Mining Applications
• Case study: Personalized Medicine
• Case Study: Drug Repositioning
• Future Challenges and Summary
Thank you! | Questions?

Ping Zhang: pzechang@us.ibm.com
Lun Yang: Lun.Yang@gmail.com