Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization

6th International Workshop on Machine Learning in Systems Biology (MLSB 2012)
Basel, Switzerland

Mehmet Gönen
mehmet.gonen@aalto.fi
http://users.ics.aalto.fi/gonen/

Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Aalto University School of Science

September 9, 2012
In This Talk

- Introduction
- Materials
- Earlier Approaches
- Kernelized Bayesian Matrix Factorization
- Results
- Conclusions
Introduction
Identifying Interactions Between Drugs and Proteins

- Functions of proteins can be modulated by drugs

- Growing knowledge about chemical space of drug compounds and genomic space of target proteins
  - high-throughput chemical compound screening with biological assays
  - high-throughput experimental projects that analyze the genome

- Limited knowledge about relationship between these two spaces
  - laborious and costly experimental procedures
Introduction
Identifying Interactions Between Drugs and Proteins

- A small number of experimentally validated interactions in existing databases
  - ChEMBL (Gaulton et al., 2012), DrugBank (Knox et al., 2011), KEGG DRUG (Kanehisa et al., 2012) and SuperTarget (Hecker et al., 2012)

- Computational methods for identifying interactions between drug compounds and target proteins
  - to guide experimentalists towards new predictions
  - to provide supporting evidence for their experimental results
Introduction
Identifying Interactions Between Drugs and Proteins

Traditional methods

1. docking simulations (Cheng et al., 2007; Rarey et al., 1996)
   - requires structural information of target protein
2. ligand-based approaches (Butina et al., 2002; Byvatov et al., 2003; Keiser et al., 2007)
   - requires a significant number of known ligands for target protein
3. literature text mining (Zhu et al., 2005)
   - can not predict unknown interactions
   - suffers from nonstandard naming practices
Introduction
Identifying Interactions Between Drugs and Proteins

- Machine learning methods operate on
  1. chemical properties of drug compounds
  2. genomic properties of target proteins
  3. known interaction network

- “Similar drug compounds are likely to interact with similar target proteins”

- Similarities can be encoded using kernel functions designed for chemical compounds and protein sequences
Four important protein families from humans

1. **Enzymes (E)**: proteins that catalyze (i.e., increase the rates of) chemical reactions
2. **Ion Channels (IC)**: proteins that regulate the flow of ions across the membrane in all cells
3. **G-Protein-Coupled Receptors (GPCR)**: proteins that sense molecules outside the cell and activate inside signal transduction pathways and cellular responses
4. **Nuclear Receptors (NR)**: proteins that are responsible for sensing steroid and thyroid hormones and certain other molecules
### Materials

#### Datasets

- **Four drug–target interaction networks from Yamanishi *et al.* (2008)**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Drugs</th>
<th>Number of Proteins</th>
<th>Number of Interactions</th>
<th>Ratio of Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>445</td>
<td>664</td>
<td>2926</td>
<td>≈ 1.0%</td>
</tr>
<tr>
<td>IC</td>
<td>210</td>
<td>204</td>
<td>1476</td>
<td>≈ 3.5%</td>
</tr>
<tr>
<td>GPCR</td>
<td>223</td>
<td>95</td>
<td>635</td>
<td>≈ 3.0%</td>
</tr>
<tr>
<td>NR</td>
<td>54</td>
<td>26</td>
<td>90</td>
<td>≈ 6.5%</td>
</tr>
</tbody>
</table>

- Only experimentally validated interactions
Materials
Chemical Data

- Drug compounds

  (a) Aspirin
  ![Aspirin structure](image)
  D00109

  (b) Paracetamol
  ![Paracetamol structure](image)
  D00217

- Structural similarity between drug compounds using SIMCOMP (Hattori et al., 2003)

- Drugs are represented as graphs
A dictionary of substructures

Each drug is a set of substructures

Chemical similarity score between two drug compounds

$$s_c(d_i, d_k) = \frac{|d_i \cap d_k|}{|d_i \cup d_k|}$$
Target proteins (two enzymes affected by paracetamol)

(a) 2FDV  (b) 3E6I

Sequence similarity between target proteins using normalized Smith-Waterman score (Smith and Waterman, 1981)

Proteins are represented as amino-acid sequences
Materials
Genomic Data

- Each protein is a string from 20-letter alphabet
  MSALGVTVALLVWAAILLVSLMWRQVHSSWNLPFGPFLPPIIGNLFQLELKNIPKSFTRL
  AQRFGPVFTLYVSQRMVVMHZKAKEALLYKDEFSGRDLPAPAFHAHRDRGIIIFNNGP
  TWKDIRRFSLTTLRNYGMGKQGNESRIQREQAHFLLEALRKTQGQPFDPFLIGCAPCNVI
  ADILFRKHDYNDEKFLRLMYLFNENFHLLSTPWQLYNNFPSFLHYPGSHRKVIKNVA
  EVKEYVSERVKEHHQLDPNCPRDLTDCLLVEMEKEKHSAERLYTMGIVTVADLFFAG
  TETTSTTLRYGLLILMMKPIEEKLHEEIDRVIGPSRIPAIKDRQEMPYMDAVVHEIQRF
  ITLVPNSLPHEATRDTIFRGYLIPKGTVVPTLDVLYDNEFDPDPKPFKPEHFLNENGG
  FKYSDFYKPFSTGKRCAGEGLARMELFLLLCAILQHFNLKPLVDPKIDLSPIHIGFGC
  IPPRYKLCVIPRS

- Genomic similarity score between two target proteins
  \[ s_g(t_j, t_i) = \frac{SW(t_j, t_i)}{\sqrt{SW(t_j, t_j)SW(t_i, t_i)}} \]
Materials
Interaction Data

- $N_d$ drug compounds denoted as $X_d = \{d_1, d_2, \ldots, d_{N_d}\}$

- $N_t$ target proteins denoted as $X_t = \{t_1, t_2, \ldots, t_{N_t}\}$

- $N_d \times N_t$ matrix of known interactions between these two sets denoted as $Y$

$$y_{ij} = \begin{cases} +1 & \text{if drug compound } d_i \text{ interacts with target protein } t_j \\ -1 & \text{otherwise} \end{cases}$$
Materials
Interaction Data

Three important out-of-sample prediction scenarios

1. To find interacting proteins from $X_t$ for a new drug $d_*$
2. To find interacting drugs from $X_d$ for a new target $t_*$
3. To estimate whether a new drug $d_*$ and a new target $t_*$ are interacting with each other

Predicting unknown drug–target interactions of given network

- Some drug–target pairs are labeled as $-1$ due to missing experimental evidence but they can be interacting in reality
Earlier Approaches
Pairwise Kernel Methods

- A binary classification task between drug–target pairs using pairwise kernel functions (Jacob and Vert, 2008; Wassermann et al., 2009)

\[ k((d_i, t_j), (d_k, t_l)) = k_c(d_i, d_k)k_g(t_j, t_l) \]

- Computationally heavy due to high number of drug–target pairs
  - calculates an \( N_d N_t \times N_d N_t \) kernel matrix between object pairs \( \Rightarrow O(N_d^2 N_t^2) \) storage complexity
  - trains a kernel-based classifier using this kernel matrix \( \Rightarrow O(N_d^3 N_t^3) \) time complexity
Earlier Approaches
Bipartite Graph Inference

- Maps drug compounds and target proteins into a unified space called *pharmacological space* (Yamanishi *et al.*, 2008, 2010)

- Mapping is done by considering
  - chemical similarity between drug compounds
  - genomic similarity between target proteins

- A drug–target pair is labeled as *interacting* if distance between them in pharmacological space is less than a threshold
Earlier Approaches
Matrix Factorization Methods

- Neighborhood methods versus latent factor models

- Matrix factorization models map both users and items into a joint latent factor space of dimensionality $R$

- User–item interactions are modeled as inner products in that space

- Best-known example is recommender systems (e.g., movie recommendation)
Earlier Approaches
Matrix Factorization Methods

latent user components

latent item components

rating matrix

out-of-sample item

out-of-sample user
**Kernelized Bayesian Matrix Factorization**

Idea Behind Proposed Method

(a) Kernel-based nonlinear dimensionality reduction (Schölkopf and Smola, 2002)

(b) Matrix factorization (Srebro, 2004)

(c) Binary classification
\[ \begin{align*}
\Lambda_d & \xrightarrow{\text{A}_d} \\
K_d & \xrightarrow{\text{G}_d} \xrightarrow{\text{F}} \xrightarrow{\text{Y}} \\
\lambda_{i,s}^i & \sim \mathcal{G}(\lambda_{i,s}^i; \alpha_\lambda, \beta_\lambda) \quad \forall (i, s) \\
a_{d,s}^i | \lambda_{d,s}^i & \sim \mathcal{N}(a_{d,s}^i; 0, (\lambda_{d,s}^i)^{-1}) \quad \forall (i, s) \\
g_{d,i}^s | a_{d,s}, k_{d,i} & \sim \mathcal{N}(g_{d,i}^s; a_{d,s}^\top k_{d,i}, \sigma_g^2) \quad \forall (s, i) \\
f_j^i | g_{d,i}, g_{t,j} & \sim \mathcal{N}(f_j^i; g_{d,i}^\top g_{t,j}, 1) \quad \forall (i, j) \\
y_j^i | f_j^i & \sim \delta(f_j^i y_j^i > \nu) \quad \forall (i, j)
\end{align*} \]

- \( \mathcal{G}(\cdot; \cdot, \cdot) \Rightarrow \) Gamma distribution
- \( \mathcal{N}(\cdot; \cdot, \cdot) \Rightarrow \) Normal distribution
- \( \delta(\cdot) \Rightarrow \) Kronecker delta
Kernelized Bayesian Matrix Factorization
Inference Using Variational Approximation

- Exact inference for our probabilistic model is intractable

- Using a Gibbs sampling approach is computationally expensive (Gelfand and Smith, 1990)

- We propose a deterministic variational approximation to make inference efficient

- Variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors (Beal, 2003)
Kernelized Bayesian Matrix Factorization
Inference Using Variational Approximation

- Factorable ensemble approximation of required posterior

\[ p(\Theta, \Xi | K_d, K_t, Y) \approx q(\Theta, \Xi) = \]
\[ q(\Lambda_d) q(A_d) q(G_d) q(\Lambda_t) q(A_t) q(G_t) q(F) \]

- We can bound marginal likelihood using Jensen’s inequality

\[ \log p(Y | K_d, K_t) \geq \]
\[ E_{q(\Theta, \Xi)} [ \log p(Y, \Theta, \Xi | K_d, K_t) ] - E_{q(\Theta, \Xi)} [ \log q(\Theta, \Xi) ] \]
Kernelized Bayesian Matrix Factorization
Inference Using Variational Approximation

\[ q(\Lambda_d) = \prod_{i=1}^{N_d} \prod_{s=1}^{R} \mathcal{G}(\lambda_{d,s}^i; \alpha_\lambda + 1/2, (1/\beta_\lambda + (a_{d,s}^i)^2 / 2)^{-1}) \]

\[ q(A_d) = \prod_{s=1}^{R} \mathcal{N}(a_{d,s}; \Sigma(a_{d,s})K_d(\tilde{g}_d^s)^\top / \sigma_g^2, (\text{diag}(\tilde{\lambda}_d^s) + K_dK_d^\top / \sigma_g^2)^{-1}) \]

\[ q(G_d) = \prod_{i=1}^{N_d} \mathcal{N}(g_{d,i}; \Sigma(g_{d,i})(\tilde{A}_d^\top k_{d,i} / \sigma_g^2 + \tilde{G}_t(\tilde{f}_i)^\top), (1/\sigma_g^2 + \tilde{G}_t \tilde{G}_t^\top)^{-1}) \]

\[ q(F) = \prod_{i=1}^{N_d} \prod_{j=1}^{N_t} \mathcal{T}\mathcal{N}(f_{j}^i; \tilde{g}_{d,i}^\top \tilde{g}_{t,j}; 1, f_{j}^i y_{j}^i > \nu) \]
Kernelized Bayesian Matrix Factorization
Inference Using Variational Approximation

Complete algorithm

Require: $K_d$, $K_t$, $Y$, $R$, $\alpha_\lambda$, $\beta_\lambda$, $\sigma_g$ and $\nu$

1: Initialize $q(A_d)$, $q(A_t)$, $q(G_d)$, $q(G_t)$ and $q(F)$ randomly
2: repeat
3: Update $q(\Lambda_d)$, $q(A_d)$ and $q(G_d)$
4: Update $q(\Lambda_t)$, $q(A_t)$ and $q(G_t)$
5: Update $q(F)$
6: until convergence
7: return $q(A_d)$ and $q(A_t)$
Results

- Our proposed method *kernelized Bayesian matrix factorization with twin kernels* (KBMF2K)

- Three experimental scenarios
  1. exploratory data analysis using low-dimensional projections
  2. predicting interactions for out-of-sample drugs
  3. predicting unknown interactions of given network
Results
Exploratory Data Analysis

- By displaying low-dimensional projections on NR dataset

- Not including 10% of drugs (proteins) and their interactions to our training network
Results
Exploratory Data Analysis

- Some important observations
  1. KBMF2K successfully captures bipartite nature of given interaction networks (i.e., two disjoint node sets)
  2. Dashed lines (i.e., interactions from training network) connect nearby drugs and proteins
  3. Projections for held-out drugs (proteins) are meaningful because they are connected to nearby proteins (drugs)

- Prediction performance with just two dimensions may not be enough, but these two-dimensional figures can be used for exploratory data analysis
Results

Predicting Interactions for Out-of-Sample Drugs

- Five replications of five-fold cross-validation over drugs

- Average AUC (area under ROC curve) values over 25 replications

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Yamanishi et al. (2010)</th>
<th>KBMF2K</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>0.821</td>
<td>0.832</td>
</tr>
<tr>
<td>IC</td>
<td>0.692</td>
<td>0.799</td>
</tr>
<tr>
<td>GPCR</td>
<td>0.811</td>
<td>0.857</td>
</tr>
<tr>
<td>NR</td>
<td>0.814</td>
<td>0.824</td>
</tr>
</tbody>
</table>

- 10.7% and 4.6% improvements on IC and GPCR datasets
Results
Predicting Interactions for Out-of-Sample Drugs

- Average AUC values with changing subspace dimensionality

- $R$ can be optimized using automatic relevance determination (Neal, 1996)
Results
Predicting Unknown Interactions of Given Network

- Experimental procedure
  1. train KBMF2K with given interaction network
  2. rank noninteracting (i.e., not known to interact) drug–target pairs with respect to their interaction scores
  3. check predicted interactions manually from latest online versions of ChEMBL (Gaulton et al., 2012), DrugBank (Knox et al., 2011) and KEGG DRUG (Kanehisa et al., 2012) databases

- If we pick top five predicted interactions, 80% of predictions (16 out of 20) is reported in at least one database
Results
Predicting Unknown Interactions of Given Network

- E dataset has 2926 interacting and 292554 noninteracting (i.e., not known to interact) drug–target pairs

<table>
<thead>
<tr>
<th>Rank</th>
<th>Pair</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CD</td>
<td>D00437</td>
<td>Nifedipine (JP16/USP/INN) cytochrome P450, family 2, subfamily C, polypeptide 9</td>
</tr>
<tr>
<td>2 CDK</td>
<td>D00542</td>
<td>Halothane (JP16/USP/INN) cytochrome P450, family 2, subfamily E, polypeptide 1</td>
</tr>
<tr>
<td>3 CD</td>
<td>D00097</td>
<td>Salicylic acid (JP16/USP) prostaglandin-endoperoxide synthase 2</td>
</tr>
<tr>
<td>4</td>
<td>D00501</td>
<td>Pentoxifylline (JAN/USP/INN) phosphodiesterase 7A</td>
</tr>
<tr>
<td>5 DK</td>
<td>D00139</td>
<td>Methoxsalen (JP16/USP) cytochrome P450, family 1, subfamily A, polypeptide 1</td>
</tr>
</tbody>
</table>

C: ChEMBL, D: DrugBank and K: KEGG
Conclusions

Summary

- A novel Bayesian formulation that combines
  - kernel-based nonlinear dimensionality reduction
  - matrix factorization
  - binary classification

- First fully probabilistic formulation proposed for drug–target interaction network inference

- Empirical evidence on four drug–target interaction networks
  - chemical similarity between drug compounds
  - genomic similarity between target proteins
Conclusions

Summary

- Propose a variational approximation for efficient inference

- Matlab implementation is available at http://users.ics.aalto.fi/gonen/kbmf2k

- An interesting direction for future research is to integrate multiple similarity measures for both drugs and proteins using *multiple kernel learning* (Gönen and Alpaydın, 2011)
  - chemical descriptors for drug compounds
  - structural descriptors for target proteins
References


References II


