Aalto University
School of Science

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

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

Materials

Datasets

- 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

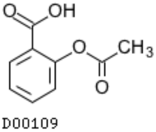
- Four drug–target interaction networks from Yamanishi *et al.* (2008)

Dataset	Number of Drugs	Number of Proteins	Number of Interactions	Ratio of Interactions
E	445	664	2926	$\approx$ 1.0%
IC	210	204	1476	$\approx$ 3.5%
GPCR	223	95	635	$\approx$ 3.0%
NR	54	26	90	$\approx$ 6.5%

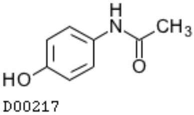
- Only experimentally validated interactions

Materials
Chemical Data

- Drug compounds



(a) Aspirin

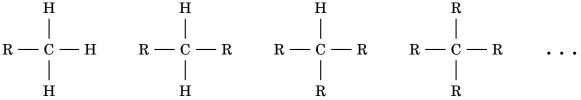


(b) Paracetamol

- Structural similarity between drug compounds using SIMCOMP (Hattori *et al.*, 2003)
- Drugs are represented as graphs

Materials
Chemical Data

- A dictionary of substructures



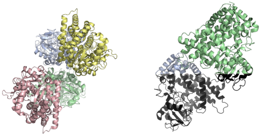
- Each drug is a set of substructures

- Chemical similarity score between two drug compounds

$$s_c(\boldsymbol{d}_i, \boldsymbol{d}_k) = \frac{|\boldsymbol{d}_i \cap \boldsymbol{d}_k|}{|\boldsymbol{d}_i \cup \boldsymbol{d}_k|}$$

Materials
Genomic Data

- 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

```
MSALGVTVALLVWAAFLLLVSMWRQVHSSWNLPPGPFPLPIIGNLFQLELKNIPKSFTRL
AQRFGPVFTLYVGSQRMVVMHGYKAVKEALLDYKDEFSGRGDLPAFHAHRDRGIIFNNGP
TWKDIRRFSLTTLRNYGMGKQGNESRIQREAHFLLEALRKTQGQPFDPTFLIGCAPCNVI
ADILFRKHFDYNDEKFLRLMYLFNENFHLLSTPWLQLYNNFPSFLHYLPGSHRKVIKNVA
EVKEYVSERVKEHHQSLDPNCPRDLTDCLLVEMEKEKHSAERLYTMDGITVTVADLFFAG
TETTSTTLRYGLLILMKYPEIEEKLHEEIDRVIGPSRIPAIKDRQEMPYMDAVVHEIQRF
ITLVPSNLPHEATRDTIFRGYLIPKGTVVVPTLDSVLYDNQEFPDPEKFKPEHFLNENGK
FKYSDYFKPFSTGKRVCAGEGLARMELFLLLCAILQHFNLKPLVDPKDIDLSPIHIGFGC
IPPRYKLCVIPRS
```

- Genomic similarity score between two target proteins

$$s_{\mathrm{g}}(\boldsymbol{t}_j, \boldsymbol{t}_l) = \frac{\mathtt{SW}(\boldsymbol{t}_j, \boldsymbol{t}_l)}{\sqrt{\mathtt{SW}(\boldsymbol{t}_j, \boldsymbol{t}_j)\mathtt{SW}(\boldsymbol{t}_l, \boldsymbol{t}_l)}}$$

Materials

Interaction Data

- N_d drug compounds denoted as $\mathbf{X}_\mathrm{d} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_{N_\mathrm{d}}\}$

- N_t target proteins denoted as $\mathbf{X}_\mathrm{t} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \dots, \boldsymbol{t}_{N_\mathrm{t}}\}$

- $N_\mathrm{d} \times N_\mathrm{t}$ matrix of known interactions between these two sets denoted as $\mathbf{Y}$

$$y_j^i = \begin{cases} +1 & \text{if drug compound } \boldsymbol{d}_i \text{ interacts with target protein } \boldsymbol{t}_j \\ -1 & \text{otherwise} \end{cases}$$

Materials
Interaction Data

- Three important out-of-sample prediction scenarios
  1. To find interacting proteins from $\mathbf{X}_{\mathrm{t}}$ for a new drug $\boldsymbol{d}_\star$
  2. To find interacting drugs from $\mathbf{X}_{\mathrm{d}}$ for a new target $\boldsymbol{t}_\star$
  3. To estimate whether a new drug $\boldsymbol{d}_\star$ and a new target $\boldsymbol{t}_\star$ 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)

  $k((\boldsymbol{d}_i, \boldsymbol{t}_j), (\boldsymbol{d}_k, \boldsymbol{t}_l)) = k_c(\boldsymbol{d}_i, \boldsymbol{d}_k) k_g(\boldsymbol{t}_j, \boldsymbol{t}_l)$

- Computationally heavy due to high number of drug–target pairs
  - calculates an $N_{\mathtt{d}} N_{\mathtt{t}} \times N_{\mathtt{d}} N_{\mathtt{t}}$ kernel matrix between object pairs $\Rightarrow \mathcal{O}(N_{\mathtt{d}}^2 N_{\mathtt{t}}^2)$ storage complexity
  - trains a kernel-based classifier using this kernel matrix $\Rightarrow \mathcal{O}(N_{\mathtt{d}}^3 N_{\mathtt{t}}^3)$ 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)

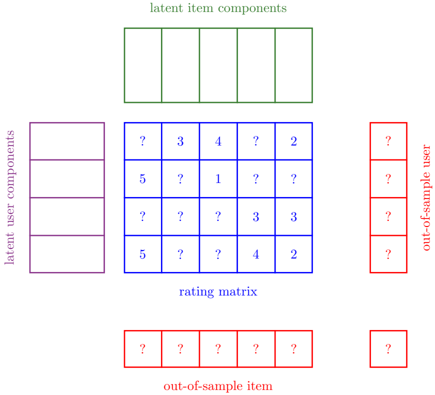
- Mapping is done by considering
  - chemical similarity between drug compounds
  - genomic similarity between target proteins

- 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 item components

latent user components

?	3	4	?	2
5	?	1	?	?
?	?	?	3	3
5	?	?	4	2

rating matrix

?
?
?
?

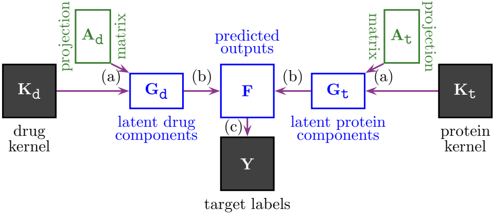
out-of-sample user

?	?	?	?	?	?

out-of-sample item

Kernelized Bayesian Matrix Factorization

Idea Behind Proposed Method



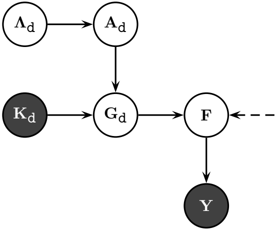
(a) Kernel-based nonlinear dimensionality reduction (Schölkopf and Smola, 2002)

(b) Matrix factorization (Srebro, 2004)

(c) Binary classification

Kernelized Bayesian Matrix Factorization

Graphical and Probabilistic Models



$$\lambda_{\mathrm{d},s}^{i} \sim \mathcal{G}(\lambda_{\mathrm{d},s}^{i}; \alpha_\lambda, \beta_\lambda) \qquad \forall(i, s)$$

$$a_{\mathrm{d},s}^{i} | \lambda_{\mathrm{d},s}^{i} \sim \mathcal{N}(a_{\mathrm{d},s}^{i}; 0, (\lambda_{\mathrm{d},s}^{i})^{-1}) \qquad \forall(i, s)$$

$$g_{\mathrm{d},i}^{s} | \boldsymbol{a}_{\mathrm{d},s}, \boldsymbol{k}_{\mathrm{d},i} \sim \mathcal{N}(g_{\mathrm{d},i}^{s}; \boldsymbol{a}_{\mathrm{d},s}^{\top}\boldsymbol{k}_{\mathrm{d},i}, \sigma_g^2) \qquad \forall(s, i)$$

$$f_j^i | \boldsymbol{g}_{\mathrm{d},i}, \boldsymbol{g}_{\mathrm{t},j} \sim \mathcal{N}(f_j^i; \boldsymbol{g}_{\mathrm{d},i}^{\top}\boldsymbol{g}_{\mathrm{t},j}, 1) \qquad \forall(i, j)$$

$$y_j^i | f_j^i \sim \delta(f_j^i y_j^i > \nu) \qquad \forall(i, j)$$

- $\mathcal{G}(\cdot; \cdot, \cdot) \Rightarrow$ 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

$$p(\boldsymbol{\Theta}, \boldsymbol{\Xi} | \mathbf{K}_{\mathrm{d}}, \mathbf{K}_{\mathrm{t}}, \mathbf{Y}) \approx q(\boldsymbol{\Theta}, \boldsymbol{\Xi}) = q(\boldsymbol{\Lambda}_{\mathrm{d}}) q(\mathbf{A}_{\mathrm{d}}) q(\mathbf{G}_{\mathrm{d}}) q(\boldsymbol{\Lambda}_{\mathrm{t}}) q(\mathbf{A}_{\mathrm{t}}) q(\mathbf{G}_{\mathrm{t}}) q(\mathbf{F})$$

- We can bound marginal likelihood using Jensen's inequality

$$\log p(\mathbf{Y} | \mathbf{K}_{\mathrm{d}}, \mathbf{K}_{\mathrm{t}}) \geq \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi} | \mathbf{K}_{\mathrm{d}}, \mathbf{K}_{\mathrm{t}})] - \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta}, \boldsymbol{\Xi})]$$

Kernelized Bayesian Matrix Factorization

Inference Using Variational Approximation

$$q(\boldsymbol{\Lambda}_{\mathrm{d}}) = \prod_{i=1}^{N_{\mathrm{d}}} \prod_{s=1}^{R} \mathcal{G}\left(\lambda^{i}_{\mathrm{d},s}; \alpha_\lambda + 1/2, (1/\beta_\lambda + \widetilde{(a^{i}_{\mathrm{d},s})^2}/2)^{-1}\right)$$

$$q(\mathbf{A}_{\mathrm{d}}) = \prod_{s=1}^{R} \mathcal{N}\left(\boldsymbol{a}_{\mathrm{d},s}; \Sigma(\boldsymbol{a}_{\mathrm{d},s})\mathbf{K}_{\mathrm{d}}\widetilde{(\boldsymbol{g}^{s}_{\mathrm{d}})^\top}/\sigma_g^2, (\mathrm{diag}(\widetilde{\boldsymbol{\lambda}^{s}_{\mathrm{d}}}) + \mathbf{K}_{\mathrm{d}}\mathbf{K}_{\mathrm{d}}^\top/\sigma_g^2)^{-1}\right)$$

$$q(\mathbf{G}_{\mathrm{d}}) = \prod_{i=1}^{N_{\mathrm{d}}} \mathcal{N}\left(\boldsymbol{g}_{\mathrm{d},i}; \Sigma(\boldsymbol{g}_{\mathrm{d},i})(\widetilde{\mathbf{A}_{\mathrm{d}}^\top}\boldsymbol{k}_{\mathrm{d},i}/\sigma_g^2 + \widetilde{\mathbf{G}_{\mathrm{t}}}\widetilde{(\boldsymbol{f}^{i})^\top}), (\mathbf{I}/\sigma_g^2 + \widetilde{\mathbf{G}_{\mathrm{t}}\mathbf{G}_{\mathrm{t}}^\top})^{-1}\right)$$

$$q(\mathbf{F}) = \prod_{i=1}^{N_{\mathrm{d}}} \prod_{j=1}^{N_{\mathrm{t}}} \mathcal{TN}\left(f^{i}_{j}; \widetilde{\boldsymbol{g}_{\mathrm{d},i}^\top}\widetilde{\boldsymbol{g}_{\mathrm{t},j}}, 1, f^{i}_{j}y^{i}_{j} > \nu\right)$$

Kernelized Bayesian Matrix Factorization

Inference Using Variational Approximation

- Complete algorithm

```
Require: $\mathbf{K}_\mathtt{d}$, $\mathbf{K}_\mathtt{t}$, $\mathbf{Y}$, $R$, $\alpha_\lambda$, $\beta_\lambda$, $\sigma_g$ and $\nu$
Initialize $q(\mathbf{A}_\mathtt{d})$, $q(\mathbf{A}_\mathtt{t})$, $q(\mathbf{G}_\mathtt{d})$, $q(\mathbf{G}_\mathtt{t})$ and $q(\mathbf{F})$ randomly
repeat
    Update $q(\mathbf{\Lambda}_\mathtt{d})$, $q(\mathbf{A}_\mathtt{d})$ and $q(\mathbf{G}_\mathtt{d})$
    Update $q(\mathbf{\Lambda}_\mathtt{t})$, $q(\mathbf{A}_\mathtt{t})$ and $q(\mathbf{G}_\mathtt{t})$
    Update $q(\mathbf{F})$
until convergence
return $q(\mathbf{A}_\mathtt{d})$ and $q(\mathbf{A}_\mathtt{t})$
```

Results

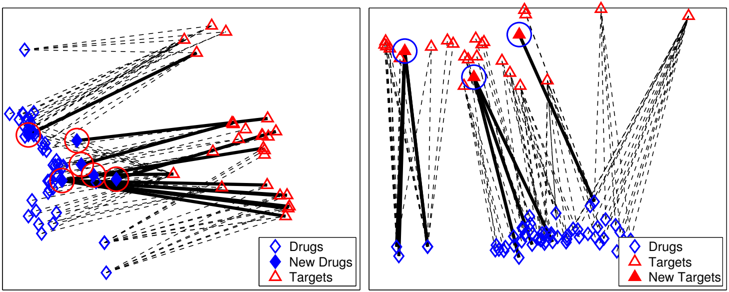
- Our proposed method *kernelized Bayesian matrix factorization with twin kernels* (KBMF2K)

- 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

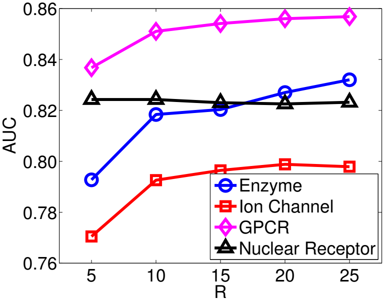
Dataset	Yamanishi *et al.* (2010)	KBMF2K
E	0.821	0.832
IC	0.692	0.799
GPCR	0.811	0.857
NR	0.814	0.824

- 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

Rank	Pair	Annotation
1 **CD**	**D00437** **1559**	**Nifedipine (JP16/USP/INN)** **cytochrome P450, family 2, subfamily C, polypeptide 9**
2 **CDK**	**D00542** **1571**	**Halothane (JP16/USP/INN)** **cytochrome P450, family 2, subfamily E, polypeptide 1**
3 **CD**	**D00097** **5743**	**Salicylic acid (JP16/USP)** **prostaglandin-endoperoxide synthase 2**
4	D00501 5150	Pentoxifylline (JAN/USP/INN) phosphodiesterase 7A
5 **DK**	**D00139** **1543**	**Methoxsalen (JP16/USP)** **cytochrome P450, family 1, subfamily A, polypeptide 1**

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

Beal,M.J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, The Gatsby Computational Neuroscience Unit, University College London.

Butina,D., Segall,M.D. and Frankcombe,K. (2002) Predicting ADME properties *in silico*: Methods and models. *Drug Discovery Today,* **7**, S83–S88.

Byvatov,E., Fechner,U., Sadowski,J. and Schneider,G. (2003) Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences,* **43**, 1882–1889.

Cheng,A.C., Coleman,R.G., Smyth,K.T., Cao,Q., Soulard,P., Caffrey,D.R., Salzberg,A.C. and Huang,E.S. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology,* **25**, 71–75.

Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. and Overington,J.P. (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research,* **40**, D1100–D1107.

Gelfand,A.E. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association,* **85**, 398–409.

Gönen,M. and Alpaydın,E. (2011) Multiple kernel learning algorithms. *Journal of Machine Learning Research,* **12**, 2211–2268.

Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society,* **125**, 11853–11865.

References II

Hecker,N., Ahmed,J., von Eichborn,J., Dunkel,M., Macha,K., Eckert,A., Gilson,M.K., Bourne,P.E. and Preissner,R. (2012) SuperTarget goes quantitative: Update on drug–target interactions. *Nucleic Acids Research,* **40**, D1113–D1117.

Jacob,L. and Vert,J.P. (2008) Protein-ligand interaction prediction: An improved chemogenomics approach. *Bioinformatics,* **24**, 2149–2156.

Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research,* **40**, D109–D114.

Keiser,M.J., Roth,B.L., Armbruster,B.N., Ernsberger,P., Irwin,J.J. and Shoichet,B.K. (2007) Relating protein pharmacology by ligand chemistry. *Nature Biotechnology,* **25**, 197–206.

Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V., Djoumbou,Y., Eisner,R., Guo,A.C. and Wishart,D.S. (2011) DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research,* **39**, D1035–D1041.

Neal,R.M. (1996) *Bayesian Learning for Neural Networks*. Springer, New York, NY.

Rarey,M., Kramer,B., Lengauer,T. and Klebe,G. (1996) A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology,* **261**, 470–489.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology,* **147**, 195–197.

References III

Srebro,N. (2004). *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology.

Wassermann,A.M., Geppert,H. and Bajorath,J. (2009) Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *Journal of Chemical Information and Modeling,* **49**, 2155–2167.

Yamanishi,Y., Araki,M., Gutteridge,A., Honda,W. and Kaneisha,M. (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics,* **24**, i232–i240.

Yamanishi,Y., Kotera,M., Kanesiha,M. and Goto,S. (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics,* **26**, i246–i254.

Zhu,S., Okuno,Y., Tsujimoto,G. and Mamitsuka,H. (2005) A probabilistic model for mining implicit ‘chemical compound-gene’ relations from literature. *Bioinformatics,* **21** (Suppl 2), ii245–ii251.