Supplementary material

DDR: Efficient computational method to predict drug–target interactions using graph mining and machine learning approaches

Rawan S. Olayan[1], Haitham Ashoor[1] and Vladimir B. Bajic[1,*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, Saudi Arabia

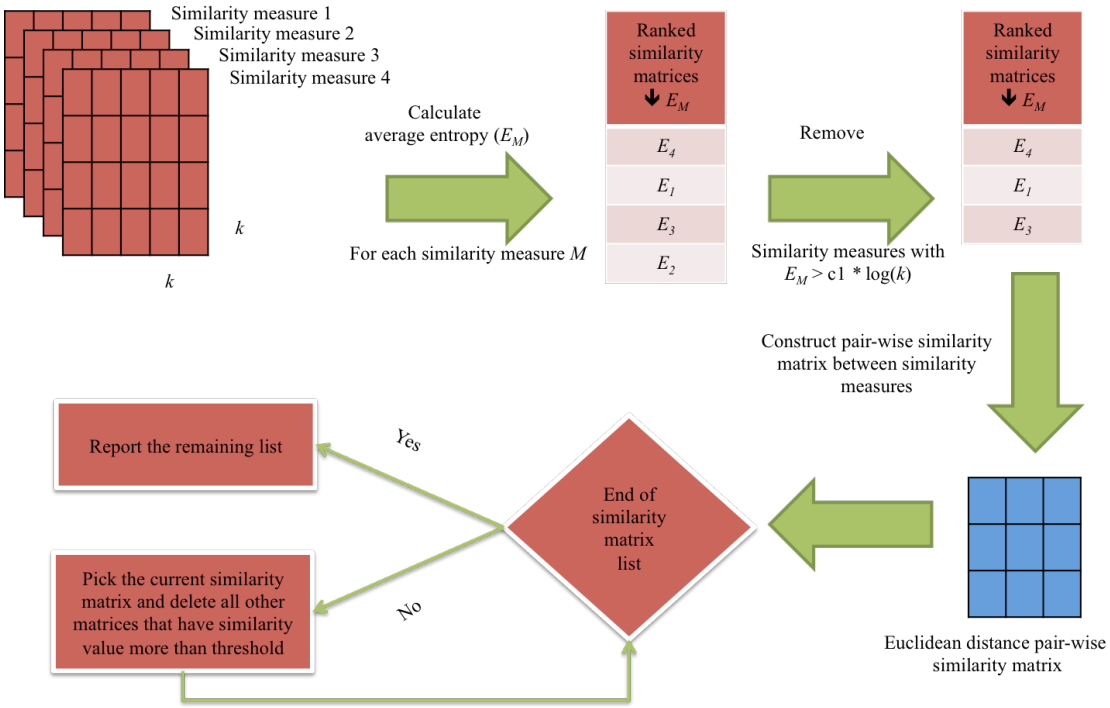
*To whom correspondence should be addressed.

Related work

In 2008, Yamanashi et al. (Yamanishi, et al., 2008) developed a statistical method based on a bipartite graph from the integration of chemical and genomic spaces to predict four classes of drug's target proteins. The dataset in (Yamanishi, et al., 2008) is considered as a golden benchmark dataset, where the target proteins are in categories of enzymes (E), ion channels (IC), G protein-coupled receptors (GPCR), and nuclear receptors (NR). Later on, several methods are proposed and demonstrated improved performance using this golden benchmark dataset. Most of existing drug-target interaction (DTI) prediction methods are designed to handle different tasks of DTI prediction such as predicting new DTIs for drugs and target proteins that have at least known DTI or predicting new DTIs for new drugs or new target proteins (Pahikkala, et al., 2015). A drug is called new if it does not have any known target proteins to interact with, and a target protein is called new if it is not targeted by any known drugs. Different recent methods utilized single type of similarity measure between drugs and single type of similarity measure between target proteins. These methods include: COSINE (Lim, et al., 2016), and NRLMF (Liu, et al., 2016). COSINE (Lim, et al., 2016) is a statistical framework that is specifically tailored to find protein targets for new chemicals with little to no available interaction data. Another class of recent DTI predictions methods aims to integrate information from different similarity measures. These methods include KronRLS-MKL (Nascimento, et al., 2016), BLM-NII (Mei, et al., 2013) and DNILMF (Hao, et al., 2017). KronRLS-MKL integrates several drug and target protein similarity measures by a linear function along with the DTI network topology for the identification of new DTIs. Their findings demonstrate that utilizing different measures of similarity between drugs and target proteins results in improved performance compared with other methods that are based on using only single similarity for drugs and single similarity for target proteins. BLM-NII infers the interaction profiles for new candidates of drugs and target proteins from interaction profiles of neighbors with strong similarities. This strategy shows its usefulness in enhancing the prediction performance for predicting DTIs of new candidates of drugs and target proteins that have no existing interactions or insufficient information in the training data. DNILMF employs a nonlinear similarity fusion technique based on similarity network fusion (SNF) method (Wang, et al., 2014) to

combine different similarity measures and then use the final diffused or combined similarity for DTI predictions. According to their model, the results based on the non-linear combination of similarity measures show better performance than other competing methods. In fact, such a prediction method that is based on non-linear integration technique of similarity measures shows better performance than other methods based on the linearly combined similarity measures (Mei, et al., 2013; van Laarhoven, et al., 2011).

Supplementary Figures



Supplementary Figure S1. Flowchart of selection process of similarity types between drugs or between target proteins, where c1 is a constant that controls how much information each similarity matrix carries; thus, c1 controls level of entropy to be selected; log(k) represents the maximum entropy value. This process selects a set of informative less-redundant set of similarities for drugs and for target proteins, separately.

Supplementary Tables

Supplementary Table S1. Summary of multiple similarity measures between drugs and between target proteins used in this study. This table shows data sources and ways to calculate different type of similarities as well as describing their importance in predicting DTIs.

Similarity and entity type:	
Type: The gene expression similarities of drugs and of target proteins. Entity: This type of similarity is calculated for each drug and target protein, separately.	Data source: Preprocessed CMap files data from (Isik, et al., 2015), where we considered only MCF7 cell line instances following the method presented in (Hizukuri, et al., 2015). Descriptor derivation: Matrix of expression profiles (comprising compounds in rows and target proteins in columns), as it explained in (Hizukuri, et al., 2015). Similarity calculation: The expression similarities of compounds and of

	target proteins, respectively, are calculated by using Pearson's correlation coefficients on the row and column profiles of the expression matrix, respectively.
	Importance: Drugs with similar expression patterns are likely to share common target proteins (Hizukuri, et al., 2015; Vilar and Hripcsak, 2016).
Type: Disease based similarity. Entity: This type of similarity is calculated for each drug and target protein, separately.	Data source: Drug-disease and target protein-disease associations are obtained from KEGG Disease (Kanehisa, et al., 2017).
	Descriptor derivation: Profile of drug–disease pairs and target protein–disease pairs that are known to be associated, each drug (or target protein) is described by a binary profile represents the presence or absence of disease name. Similarity calculation: Taminoto coefficient (TC).
	Importance: Two drugs are considered to be more similar if they have common indications, or there exist other drugs, which have common indications with them simultaneously (Dudley, et al., 2011; Luo, et al., 2016; Rodriguez-Esteban, 2016).
Type: Pathway based similarity. Entity: This type of similarity is calculated for each drug and target protein, separately.	Data source: Drug-pathway and target protein-pathway associations are obtained from KEGG Pathways.
	Descriptor derivation: Profile of drug–pathway pairs and target protein-pathway pairs that are known to be associated, each drug (or target protein) is described by a binary profile represents the presence or absence of pathway name. Similarity calculation: TC.
	Importance: Drugs acting on the same pathway may be good candidates for drug repositioning (Iwata, et al., 2017; Pan, et al., 2014; Smith, et al., 2012).
Type: Gaussian interaction profile similarity based on the topology of DTI network. Entity: This type of similarity is calculated for each drug and target protein, separately.	Data source: Known interactions between drugs and target proteins are obtained from DrugBank (Law, et al., 2014).
	It is calculated as in (van Laarhoven, et al., 2011).
	Importance: The assumption that two drugs that interact in a similar way with the target proteins in a known DTI network, will also interact in a similar way with new target proteins.
Type: Chemical structures-based molecular fingerprints similarity. Entity: Drug.	Data source: Chemical structures are obtained from DrugBank.
	Descriptor derivation: Fingerprints (i.e., CDK_Standard, CDK_Graph, CDK_Extended, CDK_Hybridization, KR, MACCS, PubChem, SIMCOMP, EC4, FC4, EC6, FC6, Lambda, Marginalized, MinMaxTanimoto, Tanimoto, and Spectrum) are generated using Kebabs (Palme, et al., 2015), Rchemcp (Klambauer, et al., 2015), Rcpi (Cao, et al., 2015), CDK (Steinbeck, et al., 2003), and SIMCOMP (Hattori, et al., 2010) tools. Similarity calculation: TC.
	Importance: Molecular fingerprints generally encode the structure of a molecule. They represent the presence or absence of particular substructures in the molecule. Calculating the similarity between two fingerprints show its practical usefulness for DTI prediction by finding matches to a substructure belonging to new molecule. The key idea behind is that chemically similar drugs tend to interact with similar target proteins (Cao, et al., 2015).
Type: Drug interactions based similarity. Entity: Drug.	Data source: DrugBank.
	Descriptor derivation: Profile of drug interactions; each drug is describing by a binary vector specifying the presence of absence of each interacting drug. Similarity calculation: TC.
	Importance: Interactive chemicals are more likely to have similar properties and thus can share similar biological functions (Chen, et al., 2012; Hu, et al., 2011; Sharan, et al., 2007; Vilar and Hripcsak, 2016).
Type: Drug side-effect based similarity. Entity: Drug.	Data source: SIDER2 (Kuhn, et al., 2016).
	Descriptor derivation: Profile of drug side-effects associations; each drug is describing by a binary vector specifying the presence of absence of each side effect keyword. Similarity calculation: TC.
	Importance: Drugs with similar target protein binding profiles tend to cause similar side effects, implying a direct correlation between target protein binding and side-effect similarity and hence a possibility to predict off-target binding (Campillos, et al., 2008; Vilar and Hripcsak, 2016).
Type: Drug ATC-code based similarity. Entity: Drug.	Data source: DrugBank.
	Descriptor derivation: Profile of drug ATC-codes associations; each

	drug is describing by a binary vector specifying the presence of absence of each ATC-code. Similarity calculation: TC and using similarity-based score from (Cheng, et al., 2013).
	Importance: Structurally similar compounds tend to have similar medical indication classes (Chen, et al., 2012; Dunkel, et al., 2008; Vilar and Hripcsak, 2016).
Type: Protein similarity based on functional annotation using gene ontology (GO). Entity: Target protein.	Data source: GOA (Barrell, et al., 2009).
	Descriptor derivation: Profile of target protein GO-terms associations, for each namespace molecular function (MF), cellular compartment (CC) and biological process (BP); each target protein is describing by a binary vector specifying the presence of absence of each GO-term. Similarity calculation: The semantic similarity is calculated using Rcpi tool.
	Importance: As structurally similar compounds tend to interact with similar biological, the functional similarity between target proteins can be established as the similarity between their GO annotation terms (Ehsani and Drablos, 2016).
Type: Protein domain based similarity. Entity: Target protein.	Data source: Pfam (Finn, et al., 2016).
	Descriptor derivation: Profile of target protein-domains associations; each target protein is describing by a binary vector specifying the presence of absence of each domain. Similarity calculation: TC.
	Importance: As structurally similar compounds tend to interact with similar biological, the similarity between target proteins can be established as the similarity between their target protein domains (Liu, et al., 2015).
Type: Protein sequence based similarity. Entity: Target protein.	Data source: UniProt, KEGG Genes.
	Similarity calculation: It is calculated using a normalized version of the Smith-Waterman (SW) algorithm (Smith and Waterman, 1981). We also calculated other sequence-based descriptors such as Mismatch and Spectrum kernels using Kebabs tool (Palme, et al., 2015).
	Importance: Recent studies follow that the target protein descriptors are as important as the compound descriptors. The key idea behind is that chemically similar drugs tend to interact with similar target proteins (Cao, et al., 2015).
Type: Proximity in protein-protein interactions (PPI) network. Entity: Target protein.	Data source: HIPPIE (Alanis-Lobato, et al., 2017).
	It is calculated as in (Perlman, et al., 2011).
	Importance: Interactive target proteins that are closer to other target proteins in the PPI network are more likely to have similar biological functions (Deng, et al., 2002).

Supplementary Table S2. Comparison results (in terms AUC scores) of DDR with the five state of the art methods (DNILMF, NRLMF, KRONRLS-MKL, COSINE, BLM-NII) using 5-repeats of 10-fold cross validation. Results are obtained under three prediction tasks (S_P, S_D, S_T) over all datasets (NR, GPCR, IC, E, DrugBank_FDA) used in this study.

Dataset	AUC obtained under prediction setting: S_P	AUC obtained under prediction setting: S_D	AUC obtained under prediction setting: S_T	Method
NR	0.92	0.90	0.88	DDR
	0.92	0.83	0.83	DNILMF
	0.93	0.88	0.83	NRLMF
	0.87	0.79	0.76	KRONRLS-MKL
		0.89		COSINE
	0.91	0.88	0.85	BLM-NII
GPCR	0.96	0.91	0.93	DDR
	0.96	0.86	0.92	DNILMF
	0.95	0.87	0.92	NRLMF
	0.91	0.81	0.84	KRONRLS-MKL
		0.88		COSINE
	0.88	0.85	0.87	BLM-NII

IC	0.98	0.94	0.97	DDR
	0.94	0.81	0.92	DNILMF
	0.98	0.80	0.93	NRLMF
	0.90	0.77	0.86	KRONRLS-MKL
		0.82		COSINE
	0.91	0.83	0.89	BLM-NII
E	0.97	0.84	0.92	DDR
	0.96	0.81	0.92	DNILMF
	0.95	0.75	0.90	NRLMF
	0.93	0.71	0.88	KRONRLS-MKL
		0.80		COSINE
	0.96	0.73	0.89	BLM-NII
DrugBank_FDA	0.96	0.91	0.86	DDR
	0.95	0.90	0.82	DNILMF
	0.93	0.89	0.80	NRLMF
	0.88	0.79	0.81	KRONRLS-MKL
		0.77		COSINE
	0.90	0.71	0.75	BLM-NII

Supplementary Table S3. Average position ranking for all methods (DDR, DNILMF, NRLMF, KRONRLS-MKL, COSINE, BLM-NII) using all datasets (NR, GPCR, IC, E, DrugBank_FDA) used in this study and under the three prediction settings (S_P, S_D, S_T).

Dataset	DDR rank	DNILMF rank	NRLMF rank	COSINE rank	KRONRLS-MKL rank	BLM-NII rank
S_P						
NR	1	3	2	NA	5	4
GPCR	1	2	3	NA	4	5
IC	1	2	5	NA	3	4
E	1	5	2	NA	3	4
DrugBank_FDA	1	2	4	NA	3	5
Average ranking	1	2.8	3.2	NA	3.6	4.4
S_D						
NR	1	4	3	2	3	5
GPCR	1	5	3	2	5	4
IC	1	4	4	3	5	2
E	1	3	4	2	6	5
DrugBank_FDA	1	3	2	6	4	5
Average ranking	1	3.8	3.2	3	4.6	4.2
S_T						
NR	1	2	4	NA	3	5
GPCR	1	2	3	NA	4	4
IC	1	2	2	NA	3	2
E	1	2	2	NA	4	3
DrugBank_FDA	1	3	2	NA	4	5
Average ranking	1	2.2	2.6	NA	3.6	3.8
Average ranking over all datasets and under three settings	1	2.933333333	3	NA	3.933333333	4.133333333

Supplementary Table S4. Comparison results (in terms of AUPR and AUC scores) of DDR with the five state of the art methods (DNILMF, NRLMF, KRONRLS-MKL, COSINE, BLM-NII) using holdout tests. Results are obtained under three prediction tasks (S_P, S_D, S_T) over DrugBank_FDA dataset used in this study.

Dataset	Prediction setting	Method	AUPR	AUC
DrugBank_FDA	S_P	DDR	0.63	0.97
		DNILMF	0.31	0.95
		NRLMF	0.34	0.93

		KRONRLS-MKL	0.32	0.93
		COSINE		
		BLM-NII	0.23	0.93
	S_D	DDR	0.42	0.92
		DNILMF	0.21	0.88
		NRLMF	0.27	0.88
		KRONRLS-MKL	0.12	0.83
		COSINE	0.1	0.88
		BLM-NII	0.09	0.87
	S_T	DDR	0.40	0.91
		DNILMF	0.12	0.86
		NRLMF	0.16	0.87
		KRONRLS-MKL	0.10	0.79
		COSINE		
		BLM-NII	0.15	0.78

Supplementary Table S5. Performance comparison (in terms of AUPR) of DDR using integrated set of selected similarity measures between drugs and between target proteins compared to combining all similarities used in this study.

Dataset	Prediction settings	AUPR (Using integrated selected set of similarities)	AUPR (Using integrated all set of similarities)
NR	S_P	0.83	0.69
	S_D	0.71	0.43
	S_T	0.64	0.42
GPCR	S_P	0.79	0.77
	S_D	0.63	0.41
	S_T	0.61	0.44
IC	S_P	0.92	0.91
	S_D	0.69	0.55
	S_T	0.80	0.70
E	S_P	0.92	0.90
	S_D	0.73	0.58
	S_T	0.82	0.75

Supplementary Table S6. The set of similarity measures selected over the five different datasets used in this study, as resulting from the similarity selection process.

Datasets	Set of selected similarity measures between drugs	Set of selected similarity measures between target proteins
NR	GIP similarity of drugs, Drug-side effect frequency-based similarity from AERS[1] database, Drugs-side effect bit-based similarity from AERS database, and drug-side effect bit-based similarity from SIDER[2] database.	GIP similarity of target proteins, and the spectrum similarity (SPEC) with k-mers length (k=4).
GPCR	GIP similarity of drugs, Drug-side effect frequency-based similarity from AERS database, Drug-side effect bit-based similarity from AERS database, and drug-side effect bit-based similarity from SIDER database.	GIP similarity of target proteins, Proximity in PPI network, and the spectrum similarity (SPEC) with k-mers length (k=4).
IC	GIP similarity of drugs, Drug-side effect frequency-	GIP similarity of target proteins, Proximity in PPI network, the spectrum similarity (SPEC) with k-mers length (k=4), and Smith-

	based similarity from AERS database, Drug-side effect bit-based similarity from AERS database, and drug-side effect bit-based similarity from SIDER database.	Waterman alignment similarity.
E	GIP similarity of drugs, Drug-side effect frequency-based similarity from AERS database, Drug-side effect bit-based similarity from AERS database, and drug-side effect bit-based similarity from SIDER database.	GIP similarity of target proteins, Proximity in PPI network, the spectrum similarity (SPEC) with k-mers length (k=4), and Smith-Waterman alignment similarity.
DrugBank_FDA	GIP similarity of drugs, drug-disease similarity, drug-pathway similarity, drug-induced gene expression similarity, and drug-ATC code similarity.	GIP similarity of target proteins, target protein-disease similarity, target protein-GO semantic similarity (namespace= BP), target protein-GO semantic similarity (namespace= CC), and target protein-GO semantic similarity (namespace= MF).

Supplementary Table S7. P-values results using label permutation test for each novel DTI in the top 5 interactions and per each dataset used in this study.

Drug ID	Drug name	Target protein ID	Target protein name	P_value
Dataset: NR				
D00348	Isotretinoin	hsa6256	RXRA	<0.01
D00585	Mifepristone	hsa2099	ESR1	0.02
D00962	Clomiphene citrate	hsa5241	PGR	0.03
D00182	Norethindrone	hsa2099	ESR1	0.02
D00951	Medroxyprogesterone acetate	hsa2099	ESR1	0.02
Dataset: GPCR				
D00049	Niacin	hsa8843	HCAR3	<0.01
D02910	Amiodarone	hsa154	ADRB2	<0.01
D02340	Loxapine	hsa1812	DRD1	<0.01
D00726	Metoclopramide	hsa1129	CHRM2	<0.01
D00674	Naratriptan hydrochloride	hsa3351	HTR1B	<0.01
Dataset: IC				
D02356	Verapamil	hsa6833	ABCC8	<0.01
D03365	Nicotine	hsa1137	CHRNA4	<0.01
D00538	Zonisamide	hsa6331	SCN5A	<0.01
D02098	Proparacaine hydrochloride	hsa8645	KCNK5	<0.01
D00775	Riluzole	hsa2898	GRIK2	<0.01
Dataset: E				
D00139	Methoxsalen	hsa1543	CYP1A1	<0.01
D00437	Nifedipine	hsa1559	CYP2C9	<0.01
D00410	Metyrapone	hsa1583	CYP11A1	<0.01
D00574	Aminoglutethimide	hsa1589	CYP21A2	<0.01
D00542	Halothane	hsa1571	CYP2E1	<0.01
Dataset: DrugBank_FDA				
DB01589	Quazepam	P47870	GABRB2	<0.01
DB00825	Menthol	P35372	OPRM1	<0.01
DB00147	Pyridoxal	P04798	CYP1A1	<0.01
DB01544	Flunitrazepam	P14867	GABRA1	<0.01
DB02546	Vorinostat	P56524	HDAC4	<0.01

Endnotes

[1] http://members.cbio.mines-paristech.fr/~yyamanishi/aers/

[2] http://sideeffects.embl.de/

References

Alanis-Lobato, G., Andrade-Navarro, M.A. and Schaefer, M.H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* 2017;45(D1):D408-D414.

Barrell, D.*, et al.* The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;37(Database issue):D396-403.

Campillos, M.*, et al.* Drug target identification using side-effect similarity. *Science* 2008;321(5886):263-266.

Cao, D.S.*, et al.* Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* 2015;31(2):279-281.

Chen, L.*, et al.* Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One* 2012;7(4):e35254.

Cheng, F.*, et al.* Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J Chem Inf Model* 2013;53(4):753-762.

Deng, M.*, et al.* Prediction of protein function using protein-protein interaction data. *Proc IEEE Comput Soc Bioinform Conf* 2002;1:197-206.

Dudley, J.T., Deshpande, T. and Butte, A.J. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 2011;12(4):303-311.

Dunkel, M.*, et al.* SuperPred: drug classification and target prediction. *Nucleic Acids Res* 2008;36(Web Server issue):W55-59.

Ehsani, R. and Drablos, F. TopoICSim: a new semantic similarity measure based on gene ontology. *BMC Bioinformatics* 2016;17(1):296.

Finn, R.D.*, et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;44(D1):D279-285.

Hattori, M.*, et al.* SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res* 2010;38(Web Server issue):W652-656.

Hizukuri, Y., Sawada, R. and Yamanishi, Y. Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. *BMC Med Genomics* 2015;8:82.

Hu, L.L.*, et al.* Predicting biological functions of compounds based on chemical-chemical interactions. *PLoS One* 2011;6(12):e29491.

Isik, Z.*, et al.* Drug target prioritization by perturbed gene expression and network information. *Sci Rep* 2015;5:17417.

Iwata, M.*, et al.* Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Sci Rep* 2017;7:40164.

Kanehisa, M.*, et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45(D1):D353-D361.
Klambauer, G.*, et al.* Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map. *Bioinformatics* 2015;31(20):3392-3394.
Kuhn, M.*, et al.* The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;44(D1):D1075-1079.
Law, V.*, et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;42(Database issue):D1091-1097.
Liu, H.*, et al.* Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 2015;31(12):i221-229.
Luo, H.*, et al.* Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;32(17):2664-2671.
Palme, J., Hochreiter, S. and Bodenhofer, U. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics* 2015;31(15):2574-2576.
Pan, Y.*, et al.* Pathway analysis for drug repositioning based on public database mining. *J Chem Inf Model* 2014;54(2):407-418.
Perlman, L.*, et al.* Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol* 2011;18(2):133-145.
Rodriguez-Esteban, R. A Drug-Centric View of Drug Development: How Drugs Spread from Disease to Disease. *PLoS Comput Biol* 2016;12(4):e1004852.
Sharan, R., Ulitsky, I. and Shamir, R. Network-based prediction of protein function. *Mol Syst Biol* 2007;3:88.
Smith, S.B.*, et al.* Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. *PLoS One* 2012;7(3):e33174.
Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* 1981;147(1):195-197.
Steinbeck, C.*, et al.* The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* 2003;43(2):493-500.
van Laarhoven, T., Nabuurs, S.B. and Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011;27(21):3036-3043.
Vilar, S. and Hripcsak, G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Brief Bioinform* 2016.