feature

Drug repurposing by integrated literature mining and drug–gene–disease triangulation

Peng Sun[1,2], Jiong Guo[3], Rainer Winnenburg[4] and Jan Baumbach[1,5], jan.baumbach@imada.sdu.dk

Drug design is expensive, time-consuming and becoming increasingly complicated. Computational approaches for inferring potentially new purposes of existing drugs, referred to as drug repositioning, play an increasingly important part in current pharmaceutical studies. Here, we first summarize recent developments in computational drug repositioning and introduce the utilized data sources. Afterwards, we introduce a new data fusion model based on n-cluster editing as a novel multi-source triangulation strategy, which was further combined with semantic literature mining. Our evaluation suggests that utilizing drug–gene–disease triangulation coupled to sophisticated text analysis is a robust approach for identifying new drug candidates for repurposing.

Introduction

The pharmaceutical industry is facing great challenges emerging from decreased speed in the discovery of new drugs and drug targets for various reasons. Although the number of approved drugs had a resurgence in 2015 [1], it was accompanied by continuously rising costs [2]. The classic conservative drug development strategy, limited to 'one drug, one target' paradigms, does not consider or evaluate the off-target effects or the probability of multiple drug indications, yet some of them have later proven successful at the market. Sildenafil and minoxidil are well-known examples. They have been repurposed for the treatment of erectile dysfunction and hair loss, respectively. Similar examples also include: ropinirole, originally developed for the treatment of Parkinson's disease but later found to be effective against restless legs syndrome [3] and potentially for selective serotonin reuptake inhibitor (SSRI)-induced sexual dysfunction [4]; and bevacizumab, originally developed to treat resistant metastatic cancers, which has been proven effective in treating abnormal retinal vascularization [5].

Drug repositioning has strong potential to provide promising solutions in current drug design. The development cycle can be reduced through repositioning by as long as 5 years compared with the traditional drug discovery pipelines [6]. Moreover, repositioned drugs have significantly reduced safety risks for patients, because almost all known drugs have been thoroughly studied with respect to their toxicity, metabolism and possible side-effects in humans [7].

Successful drug repositioning stories are rare and rather random events [7]. Well-known examples are either accidentally discovered side-effects or based on extensive research on drug properties, which is unfeasible in general and much too expensive to be applied on a large scale [8]. Thus, a major medical bioinformatics challenge is to predict high-confidence drug repositioning candidates for pharmaceutical screening, laboratory tests and clinical trials. Most existing methods for computational drug repurposing follow one of two major strategies: drug-based and disease-based approaches, depending on the data sources. They predict putative novel drug indications by exploiting detailed information of either drugs or diseases [9]. Such studies mainly focus on mining the shared properties between two drug molecules including structures [10,11] and side-effects [12]. Other methods approach the problem by computing drug–target binding properties [7] or searching for similar molecular activities [13]. The existing studies yielded fruitful and insightful results. Yet, they exclusively focus on one

http://dx.doi.org/10.1016/j.drudis.2016.10.008

aspect of drug repositioning: either the drug, the target (gene) or the disease. More-recent studies have proven the potential of combining some of the different data types by using computational information fusion [14–16]. In this review, we show the recent progress of data mining and data integration in the computational drug repositioning research, followed by a detailed description of a novel model that we suggest to integrate the information of drug, gene and disease networks using n-cluster editing.

Repositioning strategies

Methods based on drug structures

A number of publicly available databases provide a massive amount of data on molecular drug structures, chemical properties and HTS results [17–19], offering great opportunities to perform structure–property analyses useful for drug repurposing. The rationale behind this strategy is the 'structure determines properties' paradigm (i.e., molecules with similar structures tend to have similar chemical properties and, thus, act similarly on biological systems). A variety of measures based on different structural features have been used to compute the similarity of drug–molecule pairs. Such efforts include the widely used chemo- and bio-informatics library Chemical Development Kit (CDK) [20], which provides implementations for many common methods in structural chemistry and biology studies. Likewise, Swamidass [18] constructed a drug–target network based on structural similarities. A more recent trend demonstrated the benefit of integrating chemical information with other properties for computational drug repositioning. For example, Wang *et al.* [21] reported a support vector machine (SVM)-based model named PreDR implementing a customized kernel function to predict novel drug–disease associations. PreDR integrates chemical structure, molecular activity and phenotype information, such as side-effects. Similarly, Tan *et al.* [22] integrated chemical structure information (in addition to gene sequence similarities) for drug–target binding inference.

Methods based on omics data

The fast growth of omics data provides an unprecedented opportunity for computational biology to reveal more insights into drug behavior and disease mechanisms. Genome-wide expression data, in particular, are widely used to profile the effect of drug activity and have been explored for potential drug repurposing. The Connectivity Map (CMap) project by Lamb *et al.* [23] is one of the remarkable efforts aiming to construct a systematic map of the functional associations among diseases, genetic perturbation and drug behavior, based on genome-wide expression profiles of human cancer cells injected with different drugs and bioactive molecules. Thus, CMap enables systematic comparison of drug-associated gene expression profiles. For instance, Dudly *et al.* [9] followed the CMap strategy and computed a therapeutic score for every drug repositioning candidate for inflammatory bowel disease. *In vivo* model validation was performed for the most promising drug repositioning candidates. Keiser *et al.* [24] have developed a systematic tool: similarity ensemble approach (SEA), to compute the drug–target similarity by comparing the profiles of the binding ligands. Drug off-target effects have been derived and captured from the ligand-based target similarity. The top-scored repositioning candidates were later validated in an *in vivo* rodent model.

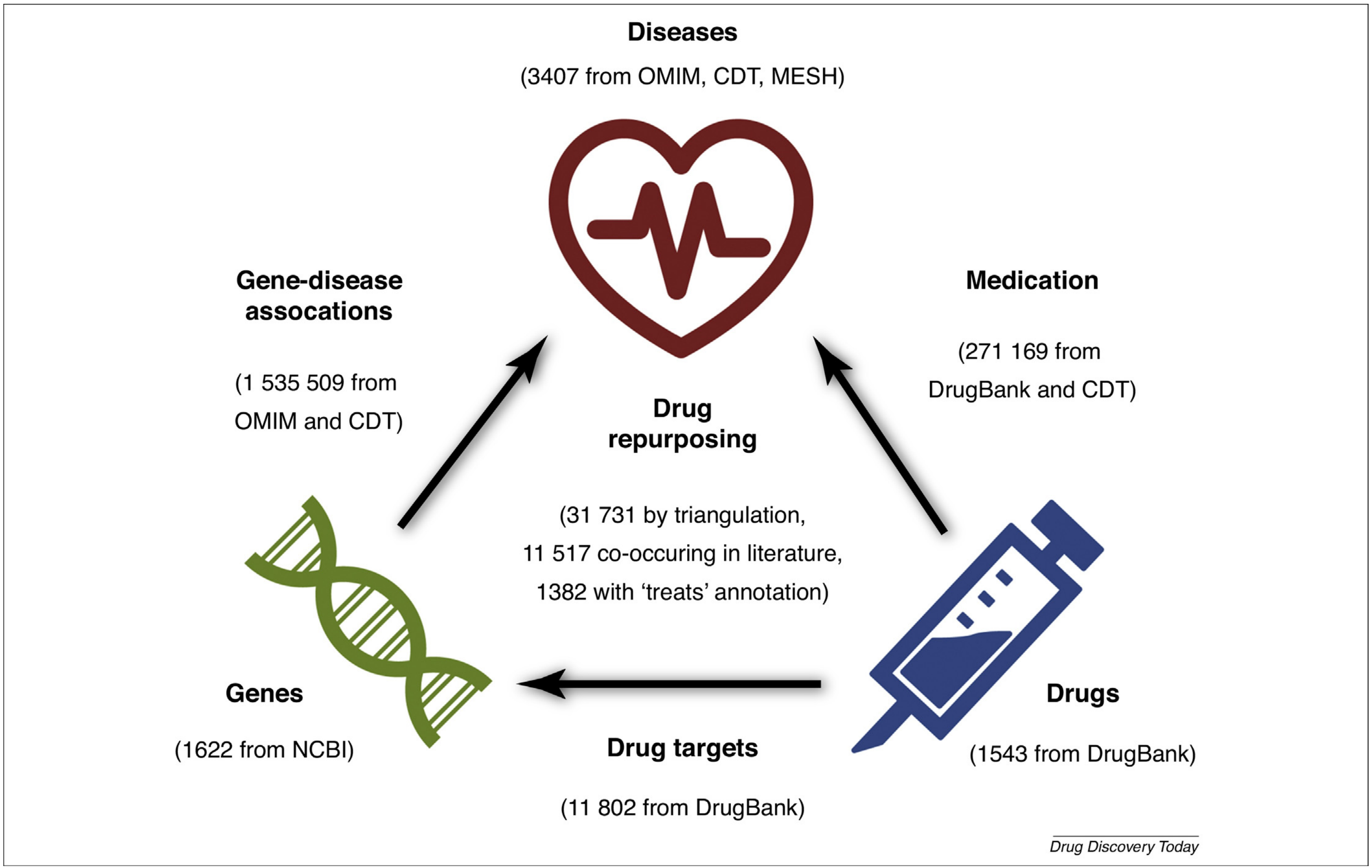
Diseases
(3407 from OMIM, CDT, MESH)
Gene-disease assocations
(1 535 509 from OMIM and CDT)
Medication
(271 169 from DrugBank and CDT)
Drug repurposing
(31 731 by triangulation, 11 517 co-occuring in literature, 1382 with 'treats' annotation)
Genes
(1622 from NCBI)
Drugs
(1543 from DrugBank)
Drug targets
(11 802 from DrugBank)
Drug Discovery Today

FIGURE 1

Overview of the input data, the main prediction principle and the results.

Methods based on phenotypes

Drug-related phenotype information also provides valuable insights to profile drug effects, subsequently supporting the discovery of new indications as indirect evidence. Ye *et al.* [25] constructed a drug–drug network from clinical side-effect information to generate putative drug–disease associations with the underlying hypothesis that shared side-effect profiles could lead to shared indications. Yang and Agarwal [26] also integrated side-effect profiles into drug repositioning features and built a naive Bayes model to suggest new drug uses. Investigating disease similarity is also a promising approach to identify drug repurposing opportunities, based on the hypothesis that similar diseases can have similar therapies. Chiang and Butte [27] derived disease similarities from shared treatments, and subsequently executed a guilt-by-association approach to predict new drug indications.

Phenome-wide association studies (PheWAS), dedicated to systematically investigating genotype-to-disease associations, have shown great potential in discovering the genetic profiles of diseases [28]. Such analyses enhance the genotype–phenotype associations detected by other studies [e.g., genome-wide association studies (GWAS)] and shed new light on drug repositioning. Rastegar-MoJarad *et al.* [15] constructed a phenotype–genotype-drug network using PheWAS data to identify multiple diseases sharing common genetic etiology that could be treatable by the same set of drugs.

Triangulation

We introduce a new model that integrates information on all relevant players (i.e., genes, drugs and diseases), and afterwards employs a semantic literature mining procedure to evaluate the findings and to increase the rate of true-positive hits. For the first part, we developed n-CluE, a novel information fusion method solving a long-standing computer science problem: weighted n-cluster editing. It addresses a special branch of graph clustering problems on n-partite graphs and can be utilized for drug repositioning by triangulating drugs (only approved drugs in DrugsBank [29] were included), genes and diseases (Fig. 1). These are connected by an edge in a graph if: (i) a drug targets a protein that is encoded by a gene; (ii) a gene is associated to a disease; or (iii) a drug is effective against a disease. This way a tripartite graph emerges which we seek to partition with minimal costs for edge modifications (i.e., insertions and deletions) and a disjoint union of tri-cliques is constructed (Fig. 2). We have developed a novel heuristic algorithm to solve this computationally challenging problem. We have implemented it into the software n-ClueE, extended it to respect confidence scores (usually *P*-values) as edge weights and applied it systematically to networks of drugs, genes and diseases (Fig. 1, Table 1). We were specifically interested in predicting novel edges between drugs and diseases, because they are candidates for drug repositioning. We call this set of edges the 'novel prediction set'. Note that we used curated and inferred gene–disease associations from CTD [30], which might impact the confidence of our predictions. The novel prediction set consists of 31 731 drug–disease pairs, which we further filtered using a co-occurrence-based literature mining procedure adopted from Rastegar-Mojarad *et al.* [15] to check for co-occurrence in at least five articles in the US National Library of Medicine bibliographic database MEDLINE. Removing non-co-occurring pairs yields a 'high confidence set' of 11 517 new drug–disease pairs. We further narrowed the literature mining to check whether the drug was explicitly mentioned to 'treat' the disease, yielding 1382 pairs, which we refer to as the 'treats annotation set'. The n-ClueE algorithm is described in detail (see supplementary material online), where we also provide an exact optimization problem definition, as well as details about the literature mining and validation

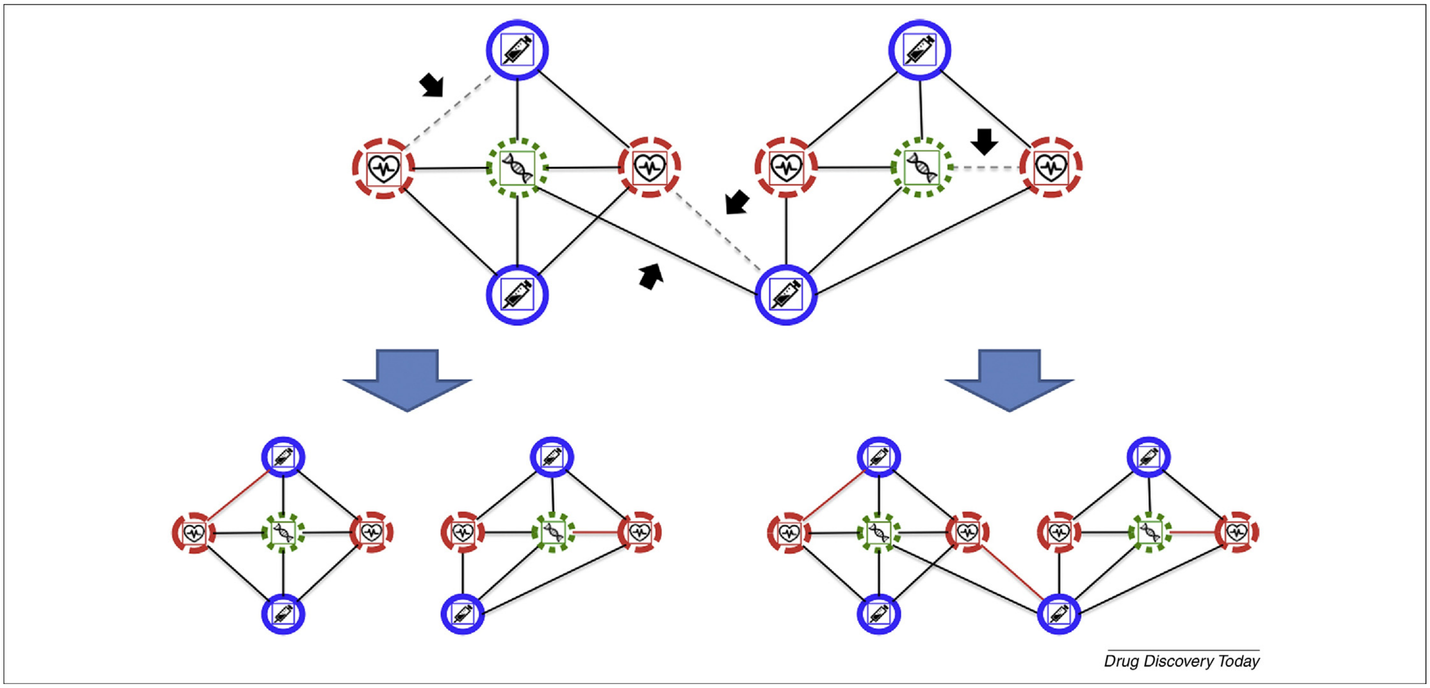

FIGURE 2

Illustration of the weighted n-cluster editing approach. The top part of the figure shows an n-partite graph constructed from drugs (blue nodes), genes (green nodes) and diseases (red nodes) using associations between them as edges. The edges possess weights that depend on the statistical significance of the corresponding associations. Whenever an edge weight falls below a certain threshold it is removed from the graph. Kept edges are visualized as solid lines connecting nodes, and some of the removed edges are visualized as gray, dashed lines. We now aim to add and delete edges from the graph such that it becomes n-transitive and minimizes an edge-modification cost function. The interesting edges are marked with black arrows. Here, depending on the concrete edge weights, either one (left figure) or two (right figure) new drug–disease pairs as well as one new gene–disease association are predicted (highlighted as red edges).

TABLE 1

Data sources used for triangulation.

Nodes and edges	Size	Source(s)
Drugs	1543	Drugbank [29]
Genes	1622	Drugbank [29], NCBI [40]
Diseases	3407	Comparative Toxicogenomics Database (CTD) [30], OMIM, MeSH [41]
Drug–gene associations	11 802	Drugbank [29]
Drug–disease associations	271 169	CTD [30], Drugbank [29]
Gene–disease associations	1 535 509	CTD [30], OMIM [42]

TABLE 2

Discovered drug repositioning examples

Drugbank ID	Drug name	Concept unique identifier	Disease name	Confidence level
DB00477	Chlorpromazine	C0041327	Tuberculosis	High
DB00859	Penicillamine	C0020542	Pulmonary hypertension	High
DB00571	Propranolol	C0677886	Ovarian epithelial cancer	High
DB01181	Ifosfamide	C2931037	Pancreatic cancer, adult	High
DB00762	Irinotecan	C2931037	Pancreatic cancer, adult	Novel
DB00635	Prednisone	C0030567	Parkinson's disease	Novel
DB04942	Tamibarotene	C1863051	Alzheimer's disease type 2	Novel

Four have literature support (confidence level high) whereas three are novel predictions. Complete lists for confidence levels in Tables S1 and S2 (see supplementary material online). Some of them have enhanced literature support (reported to 'treat' the disease) (see Table S3 in supplementary material online).

pipeline. The software implementation of n-CluE and a tutorial are available (http://nclue.compbio.sdu.dk). We provide all results sets ('novel', 'high' and 'treats' confidence) (see Tables S1–3 in supplementary material online). n-CluE triangulation suggests the drug molecule chlorpromazine (Drugbank ID: DB00477) to be associated with tuberculosis (Concept Unique Identifier: C0041327). We found this to co-occur in the literature several times (hence the high confidence level). Chlorpromazine has long been used for the therapy of psychotic disorders such as schizophrenia [31]. Our pipeline indicates an additional purpose for chlorpromazine, which is supported by several scientific articles. In the review by Zhang *et al.* [32], it is suggested that the antibacterial properties of chlorpromazine could be used for an antitubercular purpose. Another review discussed the drug resistance of pathogenic bacteria and suggested chlorpromazine to be promising as an effective antitubercular compound [33]. Likewise, n-CluE triangulation suggests the drug molecule dasatinib (Drugbank ID: DB01254) to be associated with thyroid cancer (Concept Unique Identifier: C0238463). We found this to co-occur in the literature several times (hence the high confidence level). Dasatinib is an oral Src family kinase inhibitor approved by the FDA for the treatment of lymphoblastic leukemia and chronic myelogenous leukemia [34]. *In vitro* and *in vivo* experiments demonstrated the efficacy of dasatinib controlling the growth of thyroid cancer by inhibiting the Src family kinases, which are upregulated in thyroid cancer cells [35]. An additional example for literature-supported repositioning is penicillamine (Drugbank ID: DB00859) and pulmonary hypertension (Concept Unique Identifier: C0020542), which was suggested by Oroszlán *et al.* [36]. Additionally, Xu *et al.* [37] reported that idiopathic pulmonary arterial hypertension is related to low levels of vasodilator nitric oxide (NO), and that molecules like *S*-nitroso-*N*-acetyl-D,L-penicillamine (SNAP) that provide NO in biochemical reactions can serve as a treatment. Additional (*in vitro*) experiments by Xu *et al.* gave further evidence of potential effectiveness of SNAP as a NO donor [37]. A further 11 000 additional such candidates are provided in Table S2 (see supplementary material online) for further laboratory validations and clinical studies. Research groups studying tuberculosis, for instance, will find ten interesting repositioning records, including those related to thalidomide, which has been suggested as an adjuvant treatment for tuberculosis [38]. Likewise, for pancreatic cancer investigators, we have identified 33 drug candidates with direct literature support, including salbutamol, ifosfamide, capecitabine and phenylephrine. Over 1300 more such candidates with enhanced literature support (i.e., indicating 'treatment' explicitly; see supplementary material online) can be found in Table S3 (see supplementary material online). Note that neither of the prediction sets has been filtered for potential side-effects (Table 2).

Concluding remarks

We developed the first tri-cluster editing approach, applied it to drug-disease-gene triangulation, integrated it with a literature mining pipeline and applied it to several databases for computational drug repurposing yielding over 30 000 new tricks for known drugs of which approximately 11 000 significantly co-occur in literature and over 1300 have a semantic 'treats' annotation. The utilized n-CluE algorithm solves the longstanding weighted n-cluster graph editing computer science problem. A side-effect filter based on according databases, such as SIDER [39], could further strengthen the confidence of our predictions. We anticipate that our methodology will be applied to other biomedical data processing problems in the future. In addition, we believe that our repositioning lists provide hot candidates for future screening efforts and will prove highly useful as a starting point for future clinical trials.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.drudis.2016.10.008.

References

1 Mullard, A. (2016) 2015 FDA drug approvals. *Nat. Rev. Drug Discov.* 15, 73–76

2 Plenge, R.M. (2016) Disciplined approach to drug discovery and early development. *Sci. Transl. Med.* 8 349ps15-349ps15

3 Ondo, W. (1999) Ropinirole for restless legs syndrome. *Movement Disorders* 14, 138–140

4 Worthington, J. *et al.* (2002) Ropinirole for antidepressant-induced sexual dysfunction. *Int. Clin. Psychopharmacol.* 17, 307–310

5 Rich, R.M. *et al.* (2006) Short-term safety and efficacy of intravitreal bevacizumab (Avastin) for neovascular age-related macular degeneration. *Retina* 26, 495–511

6 Elvidge, S. (2010) Getting the drug repositioning genie out of the bottle. *Life Science Leader*

7 Dudley, J.T. *et al.* (2011) Exploiting drug–disease relationships for computational drug repositioning. *Brief. Bioinform.* 12, 303–311

8 Soignet, S.L. *et al.* (1998) Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. *N. Eng. J. Med.* 339, 1341–1348

9 Dudley, J.T. *et al.* (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3, 96ra76

10 Ha, S. *et al.* (2008) IDMap: facilitating the detection of potential leads with therapeutic targets. *Bioinformatics* 24, 1413–1415

11 Gottlieb, A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496

12 Von Eichborn, J. *et al.* (2011) PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.* 39 (Suppl. 1), D1060–D1066

13 Iorio, F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14621–14626

14 Daminelli, S. *et al.* (2012) Drug repositioning through incomplete bi-cliques in an integrated drug–target–disease network. *Integr. Biol.* 4, 778–788

15 Rastegar-Mojarad, M. *et al.* (2015) Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol.* 33, 342–345

16 Mullen, J. *et al.* (2016) Mining integrated semantic networks for drug repositioning opportunities. *PeerJ* 4, e1558

17 Shi, X-N. *et al.* (2015) In silico identification and *in vitro* and *in vivo* validation of anti-psychotic drug fluspirilene as a potential CDK2 inhibitor and a candidate anti-cancer drug. *PLoS One* 10, e0132072

18 Swamidass, S.J. (2011) Mining small-molecule screens to repurpose drugs. *Brief. Bioinform.* 12, 327–335

19 Novick, P.A. *et al.* (2013) SWEETLEAD: an *in silico* database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS One* 8, e79568

20 Steinbeck, C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comp. Sci.* 43, 493–500

21 Wang, Y. *et al.* (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 8, e78518

22 Tan, F. *et al.* (2014) Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity. *Mol. BioSyst.* 10, 1126–1138

23 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935

24 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181

25 Ye, H. *et al.* (2014) Construction of drug network based on side effects and its application for drug repositioning. *PLoS One* 9, e87864

26 Yang, L. and Agarwal, P. (2011) Systematic drug repositioning based on clinical side-effects. *PLoS One* 6, e28025

27 Chiang, A.P. and Butte, A.J. (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* 86, 507

28 Hebbring, S.J. (2014) The challenges, advantages and future of phenome-wide association studies. *Immunology* 141, 157–165

29 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 34 (Suppl. 1), D668–D672

30 Davis, A.P. *et al.* (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.* 43, D914–D920

31 Crowle, A. *et al.* (1992) Chlorpromazine: a drug potentially useful for treating mycobacterial infections. *Chemotherapy* 38, 410–419

32 Zhang, Y. *et al.* (2006) New drug candidates and therapeutic targets for tuberculosis therapy. *Drug Discov. Today* 11, 21–27

33 Amaral, L. *et al.* (2007) Enhanced killing of intracellular multidrug-resistant *Mycobacterium tuberculosis* by compounds that affect the activity of efflux pumps. *J. Antimicrob. Chemother.* 59, 1237–1246

34 Chan, C.M. *et al.* (2012) Targeted inhibition of Src kinase with dasatinib blocks thyroid cancer growth and metastasis. *Clin. Cancer Res.* 18, 3580–3591

35 Chan, D. *et al.* (2012) Effect of dasatinib against thyroid cancer cell lines *in vitro* and a xenograft model *in vivo*. *Oncol. Lett.* 3, 807–815

36 Oroszlán, G. *et al.* (1992) D-penicillamine: old drug, new indication? D-penicillamine reduced pulmonary hypertension induced by free radicals. *Orvosi. Hetilap.* 133, 2835–2836 2839

37 Xu, W. *et al.* (2007) Alterations of cellular bioenergetics in pulmonary artery endothelial cells. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1342–1347

38 Fu, L. and Fu-Liu, C. (2002) Thalidomide and tuberculosis. *Int. J. Tuberc. Lung Dis.* 6, 569–572

39 Kuhn, M. *et al.* (2015) The SIDER database of drugs and side effects. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gkv1075

40 Benson, D.A. *et al.* (2015) GenBank. *Nucleic Acids Res.* 43, D30

41 Lipscomb, C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* 88, 265

42 Amberger, J.S. *et al.* (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798

Peng Sun[1,2]
Jiong Guo[3]
Rainer Winnenburg[4]
Jan Baumbach[1,5,*]

[1] *Max-Planck Institute for Informatics, Saarbrücken, Germany*
[2] *Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarbrücken, Germany*
[3] *School of Computer Science and Technology, ShanDong University, Qingdao, China*
[4] *Stanford Center for Biomedical Research, Stanford University, Stanford, CA, USA*
[5] *Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark*

*Corresponding author:.