Journal of the American Medical Informatics Association, 24(3), 2017, 614–618
doi: 10.1093/jamia/ocw142
Advance Access Publication Date: 27 September 2016
Brief Communication

Brief Communication

MeSHDD: Literature-based drug-drug similarity for drug repositioning

Adam S Brown and Chirag J Patel

Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Corresponding Author: Chirag J Patel, Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St, Boston, MA 02115, USA. E-mail: Chirag_Patel@hms.harvard.edu; Tel: (617) 432 1195.

Received 28 April 2016; Revised 17 August 2016; Accepted 23 August 2016

ABSTRACT

Objective: Drug repositioning is a promising methodology for reducing the cost and duration of the drug discovery pipeline. We sought to develop a computational repositioning method leveraging annotations in the literature, such as Medical Subject Heading (MeSH) terms.
Methods: We developed software to determine significantly co-occurring drug-MeSH term pairs and a method to estimate pair-wise literature-derived distances between drugs.
Results We found that literature-based drug-drug similarities predicted the number of shared indications across drug-drug pairs. Clustering drugs based on their similarity revealed both known and novel drug indications. We demonstrate the utility of our approach by generating repositioning hypotheses for the commonly used diabetes drug metformin.
Conclusion: Our study demonstrates that literature-derived similarity is useful for identifying potential repositioning opportunities. We provided open-source code and deployed a free-to-use, interactive application to explore our database of similarity-based drug clusters (available at http://apps.chiragjpgroup.org/MeSHDD/).

Key words: drug repositioning, similarity, MeSH terms, PubMed, metformin

BACKGROUND

Computational drug repositioning is an attractive methodology for academia and industry alike, because such a method can quickly and inexpensively nominate compounds for new indications.[1–3] Especially promising are methods that predict novel indications for currently approved drugs, as such drugs that have a substantially reduced risk of side effects. Previous approaches have generally focused on molecular evidence for repositioning, such as network studies using genomic,[4] transcriptomic,[5–8] or proteomic level information,[9] or some combination thereof.[10] Recently, however, a number of methods have been developed to utilize large-scale biomedical data from indirect sources, such as side effect profiles[11,12] and medical records data.[13]

Another source of information on approved drugs is the medical literature. In contrast to repositioning methods developed by our group and others that leverage specific types of evidence, such as differential gene expression,[5–8] methods that rely on Medical Subject Heading (MeSH) terms integrate the full spectrum of biomedical evidence, including structural, genetic, and clinical studies. The foremost repository of curated medical literature is MEDLINE®, which contains manually annotated MeSH terms for over 20 million biomedical articles. Mining this resource for drug-related information is a natural direction for computational drug repositioning, as it represents a simplified review of the literature surrounding a given drug. However, despite the fact that these data are readily and freely available, only a handful of methods have leveraged MEDLINE for repositioning. Currently available MeSH-based methods have focused on building networks connecting drugs to genes,[14,15] drugs to diseases,[16–19] or drugs to other compounds that interact when coprescribed,[20,21] but not on investigating the MeSH terms shared between drugs.

Drug-drug similarity studies are driven by the hypothesis that similar drugs should be similar in mechanism of action and be useful

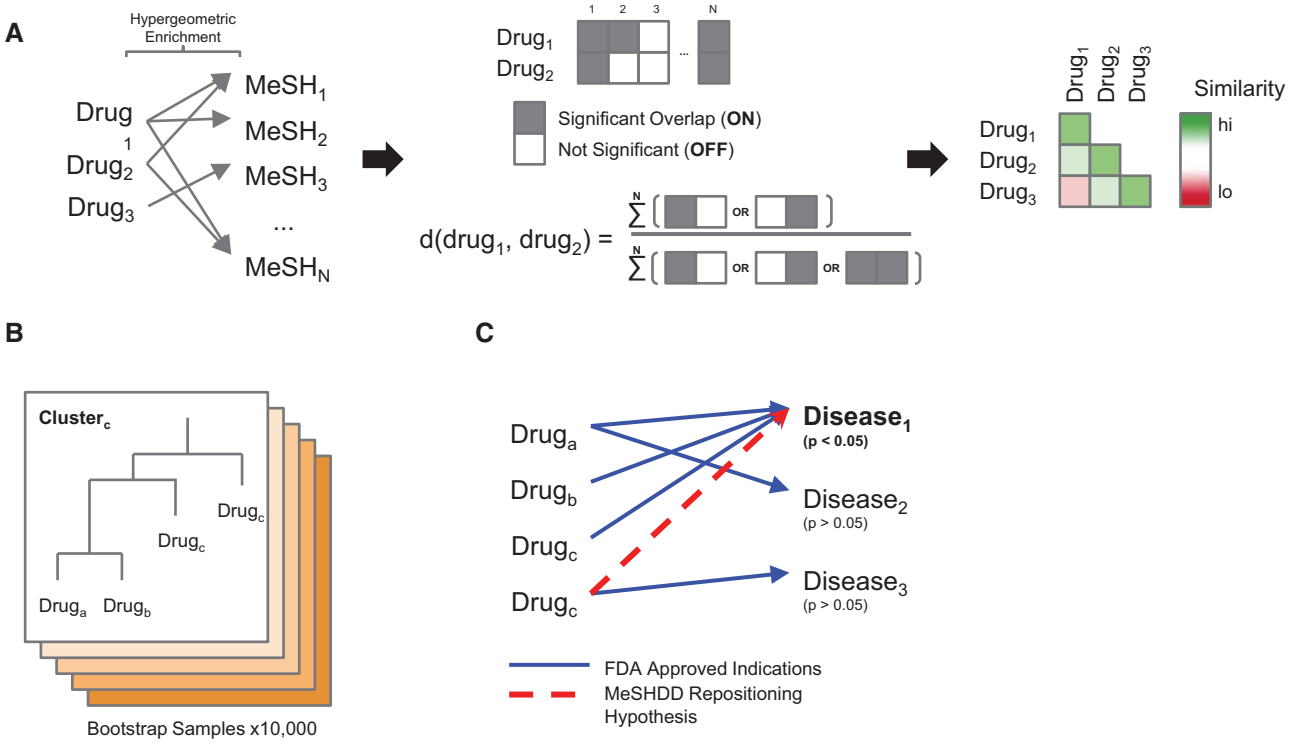

Figure 1. MeSHDD leverages literature similarity to pair drugs and diseases. (**A**) Literature similarity is assessed by calculating the bit-wise distance between 2 drugs using their significantly associated MeSH terms. (**B**) Robust clusters are defined from pair-wise distances using bootstrapping with 10 000 resamples. (**C**) Repositioning hypotheses are developed by connecting drugs to new, significantly enriched indications.

in treating a similar constellation of diseases. Drug-drug similarity has been widely applied to a variety of direct and indirect sources of evidence and with high predictive power in discovering validated repositioning opportunities.[9,12,22–24] Building on these successes, we developed MeSHDD, a MeSH-based drug-drug similarity method for computational drug repositioning.

Here, we extend methods for chemical-wise MeSH term enrichment[25] and cluster drugs based on their pair-wise similarities. We develop a methodology for predicting novel indications within drug clusters based on cluster-wise disease enrichment. We examine MeSHDD as a tool for generating repositioning hypotheses, taking metformin as a case study. Finally, we provide fully commented source code (at http://github.com/adam-sam-brown/) as well as an interactive online tool to aid investigators in generating repositioning hypotheses (at http://apps.chiragjpgroup.org/MeSHDD/).

METHODS

Drug-MeSH term overlap database construction

To identify drug-MeSH term overlap, we downloaded the main headings and corresponding chemical items file (which tracks articles referring to specific chemicals) from the MEDLINE baseline repository (accessed January 18, 2016; https://mbr.nlm.nih.gov/Download/). In parallel, we downloaded the list of 2214 US Food and Drug Administration–approved drugs from DrugBank (accessed January 18, 2015; http://www.drugbank.ca/).[26] DrugBank includes manually curated information on approved, investigational, and illicit drugs and their targets, mechanisms of action, and indications. To ensure a high degree of specificity in our Drug-MeSH term overlap, we chose to keep those MEDLINE chemicals with a case-insensitive full-length match to a DrugBank-approved drug name, resulting in 1629 overlapping drugs.

Enriched drug-MeSH terms

Using the drug-MeSH term overlap database, we calculated the enrichment for co-occurrence between each drug and MeSH term (Figure 1A).[25] To do so, we calculated a hypergeometric *P* value using the *phyper* function in the R programming language,[27] which corresponds to the probability of having as many or more drug-MeSH co-occurrences conditioned on the full set of drug-MeSH pairs. To control for multiple testing, we applied the Bonferroni correction using the *p.adjust* function in R. All associations with a Bonferroni-adjusted $P < .05$ were considered significant.

Drug cluster definition

To cluster the 1629 drugs, we leveraged a binary distance measure as implemented in the *dist* function in R. We first converted significant *P* values to binary bits, where significant entries were considered "on" (a value of 1) and nonsignificant terms were considered "off" (a value of 0). The binary distance between any 2 drugs could then be calculated as the proportion of bits for which *only 1* drug was "on" among those where *at least 1* was "on" (see Figure 1A). Highly similar drugs (and those on the diagonal) have distances close to 0, while those that are dissimilar have values close to 1. Drugs were then clustered using pair-wise distances and bootstrapped means clustering as implemented in the *clusterboot* function from the *fpc* package in R (Figure 1B).[28] We used *clusterboot* because it is optimized for large datasets and produces disjoint clusters containing all the drugs in our database. We examined a broad range of potential numbers of clusters (*k* clusters between 10 and 50)

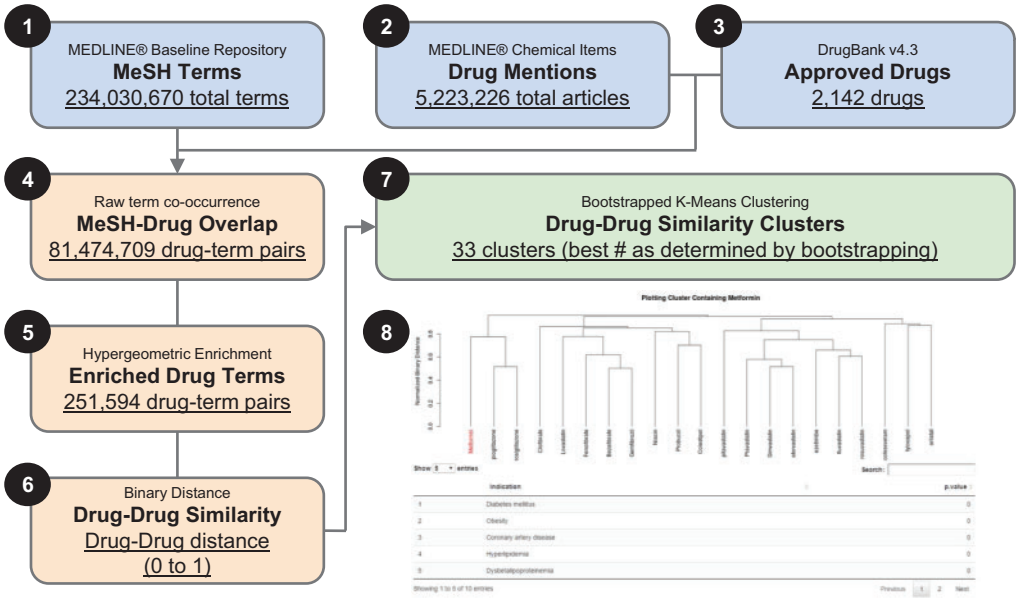

Figure 2. MeSHDD workflow for drug repositioning using MeSH terms. (1) MeSH terms are downloaded from the MEDLINE® baseline repository (2013 summary for this study). (2) Drug mentions are downloaded from the MEDLINE baseline repository, using the Chemical Items feature. (3) A list of approved drugs is downloaded from DrugBank. (4) The overlap between approved drugs and all MEDLINE MeSH terms is computed. (5) Each drug-term pair is tested for significance using the hypergeometric test for enrichment. *P* values from the test are corrected using the Bonferroni multiple-hypothesis testing method. (6) Drug-drug similarity is measured by binary distance (see Methods section). (7) Drug-drug network neighborhoods are defined using bootstrapped *k*-means, with the optimal number of clusters determined by highest mean Jaccard index. Enrichment for indications is calculated using the hypergeometric test for enrichment. (8) Screenshot from the R Shiny application, showing cluster containing metformin (used in the case study, see Results section). Height of cladogram is normalized distance between cluster members.

corresponding to a large window around the commonly used "rule-of-thumb" value for k.[29] For each value of k, we performed 100 *clusterboot* bootstraps, as recommended (see the *fpc* package manual, https://cran.r-project.org/web/packages/fpc/fpc.pdf). Goodness of clustering was assessed using the Jaccard index;[30,31] the value of k that maximized the mean Jaccard index was chosen as the optimal k (see Supplementary Figure S1). Following k selection, we performed 10 000 bootstraps to define robust clusters.

Cluster-based repositioning

To identify what indications were enriched in the putative indications in the drug-drug similarity clusters, we downloaded the Therapeutics Target Database (TTD, accessed January 23, 2016; http://bidd.nus.edu.sg/group/cjttd/).[32] The TTD contains a variety of manually curated information on over 30 000 approved and investigational drugs, including drug-disease indication information. As before, we selected only those TTD drugs with a case-insensitive full-length match to a DrugBank-approved drug name, and further restricted ourselves to those drugs tracked by MEDLINE, resulting in 1426 unique FDA-approved drugs. We then calculated the statistical overrepresentation of each disease in a given cluster using the *phyper* function in R. The resulting *P* value corresponds to the probability that a given cluster is enriched for drugs that treat a given disease, conditioned on the full set of disease-drug pairs. Bonferroni correction was applied within each cluster and across clusters to correct for multiple testing. Following correction, all *P* values that remained $< .05$ were considered significant. To assess whether more similar drugs according to our methodology would share more TTD indications, we performed ordinal logit regression using the *ordinal* package in r. Ordinal logit regression accounts for the fact that the number of shared indications between 2 drugs is an ordinal, rather than continuous. We regressed the binary distance between each pair of drugs on the number of shared indications from the TTD and assessed significance using a *P* value cutoff of .05.

RESULTS

Characteristics of data sources

We downloaded the MEDLINE baseline repository for 2013, which contained 234 030 670 total MeSH term-article pairs for 20 275 470 unique indexed PubMed articles. From this database, we extracted 81 474 709 drug-MeSH co-occurrences corresponding to 1629 unique FDA-approved drugs catalogued in DrugBank.[26] We determined enriched drug-MeSH term pairs as described above, resulting in 251 594 statistically significant pairs. In parallel, we retrieved indications for FDA-approved drugs from the TTD,[32] resulting in 1924 drug-disease pairs corresponding to 1426 unique FDA-approved drugs and 622 unique diseases (summarized in Figure 2).

Drug-drug similarity is predictive of shared disease indications

As described above, we calculated the pair-wise distance between all drug-drug pairs based on overrepresented co-occurring MeSH terms and examined the relationship between distance and number of pair-wise shared indications. As expected, if similarity were predictive of shared indication, we found that binary distance is strongly negatively correlated with number of shared indications (ordinal logit regression, $\beta_{\text{distance}} \approx -21.5$, 95% confidence interval $[-21.7, -21.3]$, $P < 2.2 \times 10^{-16}$), which corresponds to a loss of roughly 2 shared indications per 10% decrease in literature-based similarity.

This suggests that high MeSH similarity is predictive of therapeutic similarity and is a potentially useful metric for repositioning.

Drug-drug similarity clustering and disease indication enrichment

By performing bootstrapped clustering, we determined that the optimal number of drug clusters was 33, producing a median cluster size of 31 drugs. We then calculated enrichment for disease indications in the 33 clusters, which yielded predicted enrichment for 482 unique diseases of the 622 diseases considered. The median number of enriched indications per cluster was 2 (interquartile range for all 33 clusters: 1–6), which compares favorably to currently available computational repositioning methods, which typically recommend hundreds or thousands of repositioning opportunities.[5,9] MeSHDD on average provides a much tighter set of testable hypotheses.

MeSHDD R Shiny application

To enable investigators to browse the results described in this study, we developed and deployed an R Shiny application. In "drug-centric" mode, the MeSHDD application allows users to select a drug from the full list of drugs we examine, and then view other drugs in the selected drug's cluster (displayed as a dendrogram) as well as cluster-enriched disease indications (Supplementary Figure S2A). In addition to drug-centric mode, the application also allows investigators to select a disease of interest and identify clusters for which that disease is enriched (Supplementary Figure S2B). The application is available at http://apps.chiragjpgroup.org/MeSHDD/.

DISCUSSION

In this study, we describe MeSHDD, a novel literature-based repositioning methodology that leverages drug-drug similarity based on MeSH term co-occurrence. We show that our similarity measure is predictive of shared indication, with less similar drugs sharing statistically fewer indications in common. To allow investigators to generate repositioning hypotheses, we clustered drugs using their pairwise similarities and calculated disease-treatment enrichment in the resulting clusters. We also provide an interactive online tool that allows users to browse the resulting repositioning suggestions, in either a drug- or disease-centric manner. Drug-centric repositioning may be useful for academic or industry groups hoping to discover new indications for a given molecule or family of molecules, while disease-centric repositioning may be useful for identifying a small number of compounds to screen for a given disease. MeSHDD therefore represents a flexible methodology for generating a variety of different types of repositioning hypotheses.

To demonstrate MeSHDD's capability in drug-centric repositioning, we attempted to reposition the antidiabetic drug metformin. We chose metformin because, in addition to being a first-line type 2 diabetes mellitus medication,[33] it is an excellent example of successful drug repositioning. In addition to diabetes, metformin has been investigated and is currently used for a number of alternate indications.[34–36]

To generate repositioning hypotheses for metformin, we first examined the MeSHDD clustering results using the MeSHDD R Shiny application. As expected, metformin clusters with other known diabetes medications, including the glitazones, pioglitazone and rosiglitazone. Furthermore, MeSHDD correctly predicts both the primary indication, diabetes mellitus, and several investigational indications for metformin, including obesity, hyperlipidemia, and hypercholesterolemia. Interestingly, the metformin cluster is also enriched for drugs that treat cystic fibrosis (CF), linking CF to metformin (the metformin cluster contains tyloxapol, a mucus-liquefying drug). This is striking, as metformin itself is not significantly associated with CF MeSH terms. Metformin is a potent activator of AMP-activated kinase, which has recently been implicated in slowing the lung and renal pathologies of CF.[37,38] Despite initial excitement over the prospect of metformin as a well-tolerated AMP-activated kinase agonist for the treatment of CF, to our knowledge it has not yet been tested for CF in a clinical setting (from clinicaltrials.gov, accessed March 3, 2016). Using MeSHDD, we were therefore able to identify a nonobvious and testable repositioning hypothesis for the use of metformin in CF therapy.

While we have discussed the potential of MeSHDD as a flexible repositioning methodology and demonstrated its utility with a case study, we do note that it has 2 main limitations. First, MeSHDD requires that a given drug be represented in the biomedical literature in order to have the potential to share similarity with other drugs; we therefore suggest that users focus on well-studied, approved drugs rather than investigational compounds (as provided in our online tool). Second, we note that MeSH term association is agnostic to the directionality of association; for example, 2 drugs could treat and cause a symptom, respectively, and yet both could still be associated with the directionless symptom MeSH term. However, we argue that this does not generally impact performance, as our similarity metric is strongly correlated with shared indication.

CONCLUSION

Here, we have described MeSHDD, a framework for computational drug repositioning using literature-derived drug-drug similarity. Critically, we claim that MeSHDD provides an alternate way of searching the biomedical corpus for novel (and existing) uses of approved drugs. We expanded previous methods using curated MeSH terms from MEDLINE to find drug-MeSH term pairs that were enriched for co-occurrence in the medical literature and developed a method for calculating pair-wise similarities between drugs. Using this methodology, we robustly clustered 1426 FDA-approved drugs and identified within-cluster repositioning opportunities. We demonstrate the utility of MeSHDD with an end-to-end case study for metformin and identify a nonobvious but supported opportunity for the treatment of cystic fibrosis. All analysis presented in this study is fully reproducible using open-source code available from GitHub; in addition, we provide free, interactive online tools to explore the full results of MeSHDD.

FUNDING

This work was supported by National Human Genome Research Institute grant T32HG002295-12, National Institute of Environmental Health Sciences grants R00 ES023504 and R21 ES025052, a gift from Agilent Technologies, and a PhRMA fellowship.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

ASB and CJP conceived of the study. ASB conducted all statistical analyses. ASB and CJP wrote the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Readhead B, Dudley J. Translational bioinformatics approaches to drug development. *Adv Wound Care*. 2013;2:470–89.
2. Shameer K, Readhead B, Dudley JT. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Curr Top Med Chem*. 2015;15:5–20.
3. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform*. Published online first: March 31, 2015, doi:10.1093/bib/bbv020.
4. Grover MP, Ballouz S, Mohanasundaram KA, *et al*. Identification of novel therapeutics for complex diseases from genome-wide association data. *BMC Med Genomics*. 2014;7 (Suppl 1):S8.
5. Lamb J, Crawford ED, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313:1929–35.
6. Sirota M, Dudley JT, Kim J, *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med*. 2011;3:96ra77.
7. Kidd BA, Wroblewska A, Boland MR, *et al*. Mapping the effects of drugs on the immune system. *Nat Biotechnol*. Published online first: November 30, 2015, doi:10.1038/nbt.3367
8. Brown AS, Kong SW, Kohane IS, *et al*. ksRepo: a generalized platform for computational drug repositioning. *BMC Bioinformatics*. 2016;17:78.
9. Huang H, Nguyen T, Ibrahim S, *et al*. DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. *BMC Bioinformatics*. 2015;16 (Suppl 13):S4.
10. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*. 2011;7:496.
11. Campillos M, Kuhn M, Gavin A-C, *et al*. Drug target identification using side-effect similarity. *Science*. 2008;321:263–66.
12. Tatonetti NP, Ye PP, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4:125ra31.
13. Ryan PB, Madigan D, Stang PE, *et al*. Medication-wide association studies. *CPT Pharmacometrics Syst Pharmacol*. 2013;2:e76.
14. Kissa M, Tsatsaronis G, Schroeder M. Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. *Methods* 2015;74:71–82.
15. Zhang R, Cairelli MJ, Fiszman M, *et al*. Exploiting Literature-derived knowledge and semantics to identify potential prostate cancer drugs. *Cancer Inform*. 2014;13:103–11.
16. Qu XA, Gudivada RC, Jegga AG, *et al*. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*. 2009;10(Suppl 5):S4.
17. Cheung WA, Ouellette BFF, Wasserman WW. Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity. *BMC Med Genomics*. 2013;6 (Suppl 2):S3.
18. Patchala J, Jegga AG. Concept Modeling-based Drug Repositioning. *AMIA Jt Summits Transl Sci Proc*. 2015;2015:222–26.
19. Xu R, Wang Q. PhenoPredict: A disease phenome-wide drug repositioning approach towards schizophrenia drug discovery. *J Biomed Inform*. 2015;56:348–55.
20. Zhang R, Adam TJ, Simon G, *et al*. Mining Biomedical Literature to Explore Interactions between Cancer Drugs and Dietary Supplements. *AMIA Jt Summits Transl Sci Proc*. 2015;2015:69–73.
21. Zhang R, Cairelli MJ, Fiszman M, *et al*. Using semantic predications to uncover drug-drug interactions in clinical data. *J Biomed Inform*. 2014;49:134–47.
22. Iwata H, Sawada R, Mizutani S, *et al*. Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data. *J Chem Inf Model*. 2015;55:446–59.
23. Sawada R, Iwata H, Mizutani S, *et al*. Target-based drug repositioning using large-scale chemical-protein interactome data. *J Chem Inf Model*. 2015;55:2717–30.
24. Shi J-Y, Yiu S-M, Li Y, *et al*. Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods*. 2015;83:98–104.
25. Cheung WA, Ouellette BFF, Wasserman WW. Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs). *BMC Bioinformatics*. 2012;13:249.
26. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34:D668–72.
27. Development Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: the R Foundation for Statistical Computing; 2011.
28. Hennig C. fpc: Flexible Procedures for Clustering. 2015. https://CRAN.R-project.org/package=fpc. Accessed February 22, 2016.
29. Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means Clustering. *Aquat Microb Ecol*. 2013;1:90–95.
30. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal*. 2007;52:258–71.
31. Hennig C. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *J Multivar Anal*. 2008;99:1154–76.
32. Yang H, Qin C, Li YH, *et al*. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res*. Published online first: November 17, 2015, doi:10.1093/nar/gkv1230.
33. An H, He L. Current understanding of metformin effect on the control of hyperglycemia in diabetes. *J Endocrinol*. 2016;228:R97–106.
34. Dai X, Wang H, Jing Z, *et al*. The effect of a dual combination of noninsulin antidiabetic drugs on lipids: a systematic review and network meta-analysis. *Curr Med Res Opin*. 2014;30:1777–86.
35. Boland CL, Harris JB, Harris KB. Pharmacological management of obesity in pediatric patients. *Ann Pharmacother*. 2015;49:220–32.
36. Hart T, Dider S, Han W, *et al*. Toward repurposing metformin as a precision anti-cancer therapy using structural systems pharmacology. *Sci Rep*. 2016;6:20441.
37. Myerburg MM, King JD Jr, Oyster NM, *et al*. AMPK agonists ameliorate sodium and fluid transport and inflammation in cystic fibrosis airway epithelial cells. *Am J Respir Cell Mol Biol*. 2010;42:676–84.
38. Takiar V, Nishio S, Seo-Mayer P, *et al*. Activating AMP-activated protein kinase (AMPK) slows renal cystogenesis. *Proc Natl Acad Sci U S A*. 2011;108:2462–67.