Drug Normalization for Cancer Therapeutic and Druggable Genome Target Discovery

Guoqian Jiang, MD, PhD, Sunghwan Sohn, PhD, Michael T. Zimmermann, PhD, Chen Wang, PhD, Hongfang Liu, PhD, Christopher G. Chute, MD, Dr.PH
Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

Heterogeneous drug data representation among different druggable genome knowledge resources and datasets delays effective cancer therapeutic target discovery within the broad scientific community. The objective of the present paper is to describe the challenges and lessons learned from our efforts in developing and evaluating a standards-based drug normalization framework targeting cancer druggable genome datasets. Our findings suggested that mechanisms need to be established to deal with spelling errors and irregularities in normalizing clinical drug data in The Cancer Genome Atlas (TCGA), whereas the annotations from NCI Thesaurus (NCIt) and PubChem are two layers of normalization that potentially bridge between the clinical phenotypes and the druggable genome knowledge for effective cancer therapeutic target discovery.

1 Introduction

The concept of druggable genome (subset of human genome expressing proteins able to bind drug-like molecules) was initially introduced by Hopkins and Groom in 2002 (1), with the belief that an assessment of the number of molecular targets that represent an opportunity for therapeutic intervention is crucial to the development of post-genomic research strategies within the pharmaceutical industry. Their analysis suggested that there are ~3,000 estimated druggable targets in the human genome, and a total of 600-1,500 exploitable drug targets that can be captured in the intersection between the druggable genome and those genes related to disease. In 2005, Russ and Lampel re-classified the druggable gene targets with the notion of druggability (ie, the presence of protein folds that favor interactions with drug-like chemical compounds) (2). They argued that many proteins are druggable according to their structure, but modulating their biological function will not provide any therapeutic benefit. They believed that the actual drug targets are the subset of druggable proteins that possess both structural and functional features of druggability and their only real validation comes with successful clinical use. In a recent study (2013), Rask-Andersen, et al (3) analyzed established and clinical trial drug targets encoded in the human genome, identifying trends in drug development based on the potentially novel targets currently being explored in clinical trials. Their results supported the notion that the “druggable genome” is composed of a finite set of targets.

With the rapid development in cancer genome sequencing studies, the concept of druggable genome is emerging as a promising tool for identifying new potential therapeutic targets for cancer patients. However, heterogeneous drug data representation among different druggable genome knowledge resources and datasets delays effective cancer therapeutic target discovery within the broader scientific community. A number of research efforts have been taken to integrate the data sources related to drugs, diseases, chemical compounds, genes and proteins, particularly through Semantic Web Resource Description Framework (RDF) technologies (4). For instance, Chem2Bio2RDF project (5) developed a single repository by aggregating data from multiple chemogenomics repositories that are cross-linked into Bio2RDF (6) and Linked Open Drug Data (LODD) (7). The project demonstrated the potential usefulness of the system with extended SPARQL (a standard RDF query language) (8) functions in cheminformatics and bioinformatics applications. Recently, Open PHACTS, a project funded under a European grant, has the goal of delivering and sustaining an open pharmacological space (OPS) to facilitate improvements in drug discovery in academia and industry (9). The project also adopts Semantic Web RDF technologies for data integration, dataset description and publishing. While these efforts did address some aspects of drug data mappings across different sources, few of them focused on clinical drug normalization that enables cancer therapeutic and druggable genome targets discovery.

The objective of the present paper is to describe the challenges and lessons learned from our efforts in developing and evaluating such a framework of drug normalization based on standard terminologies, targeting cancer druggable genome datasets.

2 Materials and Methods
2.1 Materials

2.1.1 Drug-Gene Interaction Database (DGIdb)

DGIdb is a publically available druggable genome resource (10) that organizes the genes of the druggable genome into two main classes. DGIdb integrates data from 13 primary sources. In the present study, we downloaded the DGIdb interaction dataset in the Tab-Separated Value (TSV) format from the DGIdb website (11), which contains 17,601 drug-gene interaction entries involving 5,974 unique drug names out of total 17,634 drug mentions.

2.1.2 Clinical Drug Datasets in TCGA

TCGA clinical data including clinical drug data are publicly available from TCGA Data Portal (12). In the present study, we focused on a tumor type Glioblastoma multiforme (GBM) and downloaded the clinical drug file in biotab type (13) from TCGA Data Portal, which contains clinical drug information for all GBM participants. In this file, there are 216 unique drug names out of total 1,427 drug mentions.

2.1.3 NCI Thesaurus

NCI Thesaurus (NCIt) is a reference biomedical ontology used by the National Cancer Institute (NCI) and a growing number of other systems (14). In the present study, we utilized the version 13.08d of NCIt in OWL for building a cancer drug dictionary.

2.1.4 RxNorm

RxNorm, a nomenclature for clinical drugs, is developed by the US National Library of Medicine (15). It contains the names of prescription and many nonprescription formulations approved for human use (primarily in the USA). In RxNorm, conceptually unique medication descriptions are assigned by a concept unique identifier (RxCUI). Today, RxNorm is becoming part of Meaningful Use to support the expanding functionality of health record technology.

2.1.5 MedXN – Medication Extraction and Normalization System

MedXN is an open-source medication extraction and normalization tool developed by the NLP team at Mayo Clinic (16). It is designed to extract comprehensive medication information from unstructured/semi-structured clinical notes and normalize it to the most appropriate RxNorm concept ID (RxCUI). MedXN focuses on medication normalization by mapping the comprehensive medication description to the best matching RxCUI. The MedXN system was implemented using the Apache Unstructured Information Management Architecture (UIMA) (17). In the present study, we utilized the MedXN pipeline as a component of our system to normalize the cancer-focused drug names.

2.1.6 PubChem

PubChem (18) is an open repository for mining the biological information of small molecules. Its content is organized in three linked databases: Substances, Compounds and Bioassay. In the present study, we investigated in linking the drug data in druggable genome datasets with PubChem Substance ID and Compound ID as part of our normalization framework.

2.2 Methods

We developed a framework of drug normalization for cancer therapeutic and druggable genome target discovery. In this framework, we created a cancer-focused drug dictionary based on NCIt and developed a drug normalization algorithm that annotates the drug names from cancer clinical drug data (eg, TCGA) and druggable genome knowledge resources (eg, DGIdb) with NCIt drug codes. We also propose to use the annotations of the PubChem Substance and Compound IDs for providing an additional layer of normalization for druggable genome datasets.

2.2.1 Creating a Cancer-focused Drug Dictionary From NCIt

As NCIt is represented in a well-structured RDF/OWL model, we defined a Semantic Web SPARQL query to extract the preferred labels and synonyms (or alternative labels) of NCIt drug codes that have an asserted semantic type “Pharmacologic Substance”. As the synonyms of each NCIt code are rendered in an XML format, we parsed the XML rendering and converted it into plain TSV format.

2.2.2 Normalizing Drug Names from DGIdb and TCGA Datasets

Our aim is to develop an automatic algorithm to map the drug names in DGIdb and TCGA datasets to the drug codes in the NCIt drug list as identified in the section above. Through the algorithm, each of mapped drug names will have an NCIt code assigned.

The algorithm consists of two steps. The first step is based on a simple exact-string-match algorithm. The algorithm utilizes both preferred labels and synonyms of the drug dictionary to assign the NCIt code to matched drug names in both target datasets. The second step is based on an ingredient (IN)-based semantic match algorithm. We utilized a drug normalization tool called MedXN and the RxNorm API. The underlying mechanism of the
algorithm is to assume that two drug names are the same drug if they have the same ingredient as defined in the RxNorm. The algorithm checked three types of ingredients – ingredient (IN), precise ingredient (PIN), multiple ingredient (MIN). We ran the algorithm on those drug names that did not have exact matches in the first module.

In addition, we ran the MedXN tool against those drug names without any matches, which assigns the RxCUIs to the matched drug names. In this step, we also enabled partial match functionality. The purpose here is to investigate the proportion of the unmatched drug names (ie, those that could not be mapped to NCIt codes) that can be mapped to RxNorm RxCUIs.

2.2.3 Evaluation of the normalization algorithm

We evaluated the matched results based on the exact-string-match algorithm and the IN-based semantic match algorithm respectively and analyzed the drug names without any matches. Two authors (GJ, MZ) of the paper with expertise in clinical medicine and cancer drugs participated in the evaluation. The reviewers assessed true positive (TP) indicating those correctly matched, false positive (FP) indicating those incorrectly matched and false negative (FN) indicating those incorrectly unmatched. The standard measures of precision \( P = \frac{TP}{TP+FP} \), recall \( R = \frac{TP}{TP+FN} \) and F-measure \( F = \frac{2PR}{P+R} \) are calculated for the evaluation.

2.2.4 Linking DGIdb Drugs with PubChem Substances and Compounds

We examined the sources of DGIdb drug names and retrieved the PubChem Substance ID (SID) and Compound ID (CID) from the original database files of the two sources and linked them back to those drug names in DGIdb.

3 Results

We retrieved 11,204 NCIt drug codes with their preferred names, corresponding UMLS CUIs, and 43,517 synonyms. In total, we produced a drug list of 54,721 drug names (combining preferred names and synonyms) with 11,204 unique NCIt codes, which are used for the drug normalization algorithm.

DGIdb and TCGA cancer drug data contain many duplicated drug names. We removed duplications and performed our normalization algorithm on the set of unique drug names. In total, there are 5,974 unique drug names in DGIdb and 216 drug names in TCGA GBM. Approximately 31% (1835/5974) of DGIdb drug names were mapped to NCIt codes, in which 45 matches were identified using RxNorm IN-based algorithm. 44% (96/216) of TCGA GBM drug names were mapped to NCIt codes, in which 19 matches were identified when we used a RxNorm IN-based mapping.

As multiple drug names can be mapped to a single NCIt code, we also counted the number of unique NCIt codes for the matched drug names. Table 1 provides some examples of such cases. The finding demonstrated that the synonyms of NCi drug codes are very useful in identifying the matches.

**Table 1.** The examples of multiple drug names mapped to a single NCIt code.

| dataset | drug name in dataset | UMLS CUI | NCIt Code |
|---------|----------------------|----------|-----------|
| DGIdb   | SODIUM PHENYL BUTYRATE | C0718066 | 440       |
|         | PHENYL BUTYRATE       |          |           |
|         | MS-275                | C1510480 | C1863     |
|         | ENTINOSTAT            |          |           |
|         | SNDX-275              |          |           |
|         | VATALANIB             | C0912586 | C1868     |
|         | PTK787/ZK 222584      |          |           |
| TCGA [GBM] | EMD 121974           | C0971473 | C1834     |
|         | Cilengitide           |          |           |
|         | Bevacizumab           | C0796392 | C2039     |
|         | Avastin               |          |           |
|         | BCNU                  | C0007257 | C349      |
|         | Carmustine            |          |           |
|         | Gliadel               |          |           |
|         | Carmustin             |          |           |

Table 2 shows the evaluation results on both DGIdb (for 1835 matched drug names) and TCGA GBM datasets. The evaluation demonstrated the precision was high for both datasets. The initial recall for TCGA GBM is 51.3%, which is sub-optimal. After correcting spelling mistakes, the recall increased to 98.0%, indicating a mechanism should be
developed to deal with the spelling errors/irregularities (see more discussions in the Discussion section below). We were not able to produce the recall for DGIdb in this preliminary study because 70% of drug names in DGIdb could not be mapped to NCIt codes with the current solution. Instead, we just used the MedXN tool to check the number of RxCUI matches. Out of 4,139 unmatched drug names in DGIdb, we identified 878 (21%) RxCUI matches, in which 198 RxCUIs are exact matches and 680 are partial matches.

Table 2. Evaluation results on both DGIdb (for 1835 matched drug names) and TCGA GBM datasets.

| Drug data                        | precision | recall | F-measure |
|----------------------------------|-----------|--------|-----------|
| DGIdb                            | 100%      | -      | -         |
| TCGA GBM                         | 99.0%     | 51.3%  | 67.6%     |
| TCGA GBM (after spelling correction) | 99.0%    | 98.0%  | 98.5%     |

In addition, we found that DrugBank and TTD are the two major sources that contributed most of drug names to the DGIdb while the two sources have cross-mappings to PubChem Substances and Compounds through the annotations of Substance ID (SID) and Compound ID (CID). The results indicated that more than 70% of drug names in the matched category have cross-mappings to PubChem whereas less than 50% of drug names in the Not-Matched category have the cross-mappings.

4 Discussion

The drug normalization effort of the study is motivated by increasing importance of druggable genome knowledge resources and cancer genomic datasets in identifying potential cancer therapeutic targets. In this study, we developed a drug normalization algorithm targeting on cancer-focused drugs. The algorithm is formed as a pipeline prototype comprising two modules: a simple exact-string-match algorithm and an ingredient (IN)-based semantic match algorithm. Using the exact-string-match algorithm, about 31% of DGIdb drug names are mapped to the NCIt drug codes whereas about 36% of drug names in TCGA GBM dataset had exact matches. Using the IN-based algorithm, we did identify more matches through the algorithm, with the match rate increased to 34% for DGIdb and 44% for TCGA GBM. In addition, we found that we need a spelling check module for the drug normalization of TCGA drug data because a large portion of TCGA drug names has either spelling errors or irregularities (eg, abbreviations/name combinations). We noticed that the column “drug name” in the TCGA clinical drug file is defined by a CDE “Prior Therapy Regimen Text” in the TCGA BCR data dictionary, which is a “text identification of the individual agent(s) used as part of a prior treatment regimen”. This means that the drug names are collected by free text input. For a practical solution, we can either compile a spelling error list (controlled vocabulary) that can help convert the spelling errors to standard drug labels, or utilize drug name recognition services with spelling error correction function. For example, the RxNorm API provides an Approximate Matching service for such purpose (19, 20). In the future, we plan to incorporate a spelling correction service (19, 21) as a preprocessing module into our drug normalization pipeline. However, such a solution would still require manual curation to ensure an appropriate match was made. To solve the problem in a long run, ideally, a cancer drug dictionary with standard codes should be used in the entry point for drug data collection. For example, a previous study (22) demonstrated that for text-based medication orders, the misspelling rate is around 17%, whereas the errors have been dramatically decreased after migrating to a Computerized Physician Order Entry (CPOE) system. We believe that such standard data entry applications can be powered by standards-based terminology services like LexGrid-enabled LexEVS services (23) and recent development in the standard of Common Terminology Services 2 (CTS2) (24).

We performed a preliminary evaluation for the matching results of both DGIdb and TCGA GBM datasets. For TCGA GBM dataset, we had a high precision of 99.0% and a modest recall of 51.3%. We noticed the low recall is mainly caused the spelling errors or irregularities. After manual spelling error correction, the recall increased to 98.0%. For DGIdb dataset, we also had a high precision of 100% for 1,845 matched drug names. We did not evaluate the recall because most of the 4,139 unmatched drug names are generalized organic molecule names (not standard International Union of Pure and Applied Chemistry (IUPAC) formulas). Organic molecule names are more difficult for NLP solutions to handle, given their possible degeneracy and that addition of functional groups (single word, suffix, or prefix) can drastically change the nature of the molecule described. Previous studies have demonstrated the utility of cross-referencing to PubChem for chemical structure and protein content of the resources (including ChEMBL, DrugBank, Human Metabolome Database and TTD) widely used in the pharmacogenomics arena (25). We retrieved the existing PubChem annotations for unmatched records through their parent sources (DrugBank and TTD), successfully mapping more than 50% to PubChem SID and CID annotations. Linking DGIdb
drug names to clinical drug information in PubChem will provide an additional layer of drug identification and annotation potentially useful for improving our drug normalization algorithm in future.

In summary, our drug normalization algorithm could produce high-precision results for annotating cancer-focused druggable genome drug names in NCIt codes. Our findings suggested that mechanisms need to be established to deal with spelling errors and irregularities in normalizing clinical drug data in TCGA, whereas the annotations from NCIt and PubChem are two layers of normalization that potentially bridge between the clinical phenotypes and the druggable genome knowledge for effective cancer therapeutic target discovery.

References

1 Hopkins AL, Groom CR. The druggable genome. Nature reviews Drug discovery. 2002 Sep;1(9):727-30.
2 Russ AP, Lampel S. The druggable genome: an update. Drug discovery today. 2005 Dec;10(23-24):1607-10.
3 Rask-Andersen M, Masuram S, Schioth HB. The Druggable Genome: Evaluation of Drug Targets in Clinical Trials Suggests Major Shifts in Molecular Class and Indication. Annual review of pharmacology and toxicology. 2013 Aug 30.
4 W3C RDF Standard. [cited March 5, 2014]; Available from: http://www.w3.org/RDF/
5 Chen B, Dong X, Jiao D, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics. 2010;11:255.
6 Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics. 2008 Oct;41(5):706-16.
7 Jentzsch A, Zhao J, Hassanzadeh O, Cheung K, Sanwald K, Andersson B. Linking open drug data. Proceedings of the International Conference on Semantic Systems (I-SEMANTICS'09); 2009: Graz, Austria OpenURL; 2009.
8 SPARQL query language for RDF. [cited March 5, 2014]; Available from: http://www.w3.org/TR/rdf-sparql-query/
9 Williams AJ, Harland L, Groth P, et al. Open PHACTS: semantic interoperability for drug discovery. Drug discovery today. 2012 Nov;17(21-22):1188-98.
10 Griffith M, Griffith OL, Coffman AC, et al. DGiDb: mining the druggable genome. Nature methods. 2013 Oct 13.
11 DGiDb. [cited February 27, 2017]; Available from: http://dgidb.genome.wustl.edu/downloads
12 TCGA Data Portal. [cited February 27, 2014]; Available from: https://tcga-data.nci.nih.gov/tcga/
13 TCGA GBM Clinical Drug File. [cited February 27, 2014]; Available from: https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/ber/biobat/clin/clinical_drug_gbm.txt
14 de Coronado S, Wright LW, Fragoso G, et al. The NCI Thesaurus quality assurance life cycle. Journal of biomedical informatics. 2009 Jun;42(3):530-9.
15 Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association : JAMIA. 2011 Jul-Aug;18(4):441-8.
16 Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an Open Source Medication Extraction and Normalization Tool for Clinical Text. J Am Med Inform Assoc (in press). 2014.
17 MedXN. [cited February 27, 2014]; Available from: http://sourceforge.net/projects/ohnlp/files/MedXN/
18 PubChem URL. [cited March 6, 2014]; Available from: http://pubchem.ncbi.nlm.nih.gov
19 Peters L, Kapusnik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug names. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:1117-26.
20 Approximate Matching in the RxNorm API. [cited March 3, 2014]; Available from: http://rxnav.nlm.nih.gov/RxNormApproxMatch.html
21 Sayle R, Xie PH, Muresan S. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. Journal of chemical information and modeling. 2012 Jan 23;52(1):51-62.
22 Zhou L, Plasek JM, Mahoney LM, Chang FY, DiMaggio D, Rocha RA. Mapping Partners Master Drug Dictionary to RxNorm using an NLP-based approach. Journal of biomedical informatics. 2012 Aug;45(4):626-33.
23 Pathak J, Solbrig HR, Buntrock JD, Johnson TM, Chute CG. LexGrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. Journal of the American Medical Informatics Association : JAMIA. 2009 May-Jun;16(3):305-15.
24 Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association : JAMIA. 2013 Dec;20(c2):e341-8.
25 Southan C, Sitzmann M, Muresan S. Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. Molecular informatics. 2013 Dec;32(11-12):881-97.