Chapter 14

Systematic Drug Repurposing Through Text Mining

Luis B. Tari and Jagruti H. Patel

Abstract

Drug development remains a time-consuming and highly expensive process with high attrition rates at each stage. Given the safety hurdles drugs must pass due to increased regulatory scrutiny, it is essential for pharmaceutical companies to maximize their return on investment by effectively extending drug life cycles. There have been many effective techniques, such as phenotypic screening and compound profiling, which identify new indications for existing drugs, often referred to as drug repurposing or drug repositioning. This chapter explores the use of text mining leveraging several publicly available knowledge resources and mechanism of action representations to link existing drugs to new diseases from biomedical abstracts in an attempt to generate biologically meaningful alternative drug indications.

Key words Drug repurposing, Alternative drug indications, Drug repositioning

1 Introduction

The current drug discovery and development model is perceived as a costly and time-consuming process [1]. To reduce cost and shorten the duration for drug development, drug repurposing, also known as drug repositioning, has become an attractive alternative to traditional drug development. Drug repurposing is the process of finding a new indication for existing drug compounds. In other words, it is a research process on how an existing drug can be used for disease treatment other than its original indication. Drug re-profiling is advantageous because it bypasses many expensive drug development steps, such as in vitro and in vivo screening, chemical optimization, toxicology studies, and formulation development. Consequently, financial and development risks are reduced, and the typical 10–17-year drug development process can be shortened to 3–12 years [2]. The most cited success story for drug repositioning is sildenafil, an angina treatment developed by Pfizer. During clinical trials, it was noted that patients suffering from erectile dysfunction had improvement in their conditions. Sildenafil went on to become the blockbuster drug more commonly known as Viagra®.

Vinod D. Kumar and Hannah Jane Tipney (eds.), *Biomedical Literature Mining*, Methods in Molecular Biology, vol. 1159, DOI 10.1007/978-1-4939-0709-0_14,

Further studies showed yet another therapeutic indication in treating pulmonary arterial hypertension, whereby sildenafil was marketed as Revatio®. Mechanistically, the additional indications could be explained. Sildenafil is an inhibitor of phosphodiesterase-5 (PDE-5), which is known to be expressed in pulmonary hypertensive lungs and plays a role in regulating blood flow to the penis [3].

The main concept behind drug repurposing is that novel drug indications can be identified based on the principle that the primary drug target can be associated with diseases other than its original drug indication. In addition, as drugs can act on multiple targets, secondary targets can be utilized for novel drug indications as well. Several systematic approaches for finding new uses for old drugs have been proposed. One method with much literature support leverages chemical compound similarity [4]. Since similar drug compounds have comparable target profiles, novel targets can be identified for a compound by analyzing similar compound activity. Another approach to identify alternative drug indications includes finding drugs that share a significant number of side effects [5, 6]. Drugs with similar effects may have similar actions, linking the side effects to disease. Drug *D* is proposed to be a candidate for the treatment of disease *Dise* if *D* shares side effects that are induced by a drug class currently used for *Dise* treatment [5]. Finally, gene expression signatures have been used to reposition drugs whereby a drug signature opposite to a disease signature is proposed to be a potential treatment for the disease [7]. Readers can refer to [8] for a comprehensive computational drug repurposing method review.

With the vast pharmacological and biological knowledge available in literature, finding novel drug indications using *in silico* approaches has become increasingly feasible. Literature-based discovery methods go a step further by identifying relevant knowledge through text mining so that new knowledge can be inferred from existing knowledge [9]. Swanson's ABC model [10] is a popular literature-based discovery methodology that links two concepts through a commonly shared concept. A notable finding identified from the Swanson's ABC model was the proposed use of fish oil to treat Raynaud's syndrome, which was later clinically validated [10, 11]. Scientific concepts *A* and *C* form a relationship when concept *A* co-occurs with concept *B* in one publication while concepts *B* and *C* co-occur in another publication. Variations of Swanson's ABC model have been described in the literature for indirect relationship identification [12, 13]. However, approaches based on concept co-occurrence within abstracts tend to generate too many hypotheses. Another direction for network-based approaches aims to uncover knowledge through biological networks. DrugMap Central [14] is a network-based approach that utilizes information on chemical structures, drug targets, and signaling pathways

for users to visualize and identify alternative drug indications. However, these co-occurrence and network-based approaches also generate many hypotheses, and identifying new drug indications from large networks can be time consuming.

2 Materials

A critical step in performing systematic drug repurposing is choosing appropriate knowledge sources. While there is an extensive list of publicly available databases that capture assorted biological knowledge (*see* http://www.pathguide.org/ for a list of interaction databases), most manually curated databases have poor literature coverage due to the resource-demanding curation process. More importantly, it is common for interaction databases not to capture the interaction details required for inference. For example, interaction-type descriptions are typically not included in these interaction databases. Medline is an ideal resource to obtain such detailed information on biologic interactions. For drug repurposing, Medline abstracts are utilized for obtaining gene–disease relationships that describe associations between gene expression regulation and diseases as well as protein–protein interaction relationships that capture the induction or the inhibition between protein pairs.

For the remainder of this section, we describe resources that complement the Medline knowledge source. Specifically, these resources are targeted for acquiring knowledge on cancer-related genes and drug–target interactions.

2.1 Gene Ontology

The Gene Ontology [15] is a hierarchical controlled vocabulary that includes three independent ontologies for biological processes, molecular functions, and cellular components. Standardized terms in the Gene Ontology describe gene roles and gene products in any organism. The Gene Ontology itself does not contain organism gene products. Rather, gene product biological roles are kept in Gene Ontology annotation form. For example, the Gene Ontology annotation is a useful resource in identifying genes that are associated with cancer-related biological processes. In particular, the following Gene Ontology terms and the corresponding descendants are selected as antitumor biological processes: negative regulation of cell proliferation (GO:0008285), positive regulation of apoptosis (GO:0043065), and negative regulation of angiogenesis (GO:0016525). On the other hand, tumor-promoting biological processes include these Gene Ontology terms and corresponding descendants: positive regulation of cell proliferation (GO:0008284), negative regulation of apoptosis (GO:0043066), and positive regulation of angiogenesis (GO:0045766).

2.2 UniProt

The UniProt Knowledge Base (UniProtKB) (http://www.uniprot.org/) is the largest protein sequence repository. In addition to protein sequence information, UniProtKB also includes manual annotation on proteins, and it is an ideal resource to obtain a cancer gene list. In particular, the keywords "oncogene" and "tumor suppressor" are used as search criteria with the results limited to human genes only.

2.3 NCBI Gene

The NCBI Gene (http://www.ncbi.nlm.nih.gov/gene) is a knowledge base that contains about 12 million curated gene records. Similar to UniProt, NCBI Gene is leveraged to identify cancer genes using the keywords "oncogene" and "tumor suppressor" with results restricted to human genes.

2.4 CancerQuest

CancerQuest (http://www.cancerquest.org) is a resource with information on cancer biology and treatment. The CancerQuest tool maintains a list of oncogenes (http://www.cancerquest.org/oncogene-table) and tumor suppressors (http:// www.cancerquest.org/tumor-suppressors-table).

2.5 DrugBank

Drug–target interactions are also essential for systematic drug repurposing. DrugBank [16] is a comprehensive knowledge base for drugs, drug actions, and drug targets. The drug–target interactions obtained from DrugBank include modulation definitions such as antagonist or agonist.

3 Methods

An important component behind our approach in performing systematic drug repurposing is in acquiring information relevant to drug mechanism. The semantics of the acquired biological interactions is leveraged to infer novel drug indications. By utilizing semantics and automated reasoning, we aim to produce novel drug indications that are accompanied by the biological mechanism behind the hypotheses as explanation.

Our approach, as shown in Fig. 1, can be divided into three main components: (1) the *knowledge representation component*, (2) the *knowledge acquisition component*, and (3) the *reasoning component*. In order to automatically propose alternative drug indications, it is necessary to first represent the drug mechanism in logic rule form. The knowledge acquisition component includes publicly available curated sources as well as relevant facts for drug indication identification acquired using text mining. With the facts gathered by the knowledge acquisition component and the logic rules defined in the knowledge representation component, the reasoning engine utilizes the logic rules to find interactions that link drugs with corresponding drug indications.

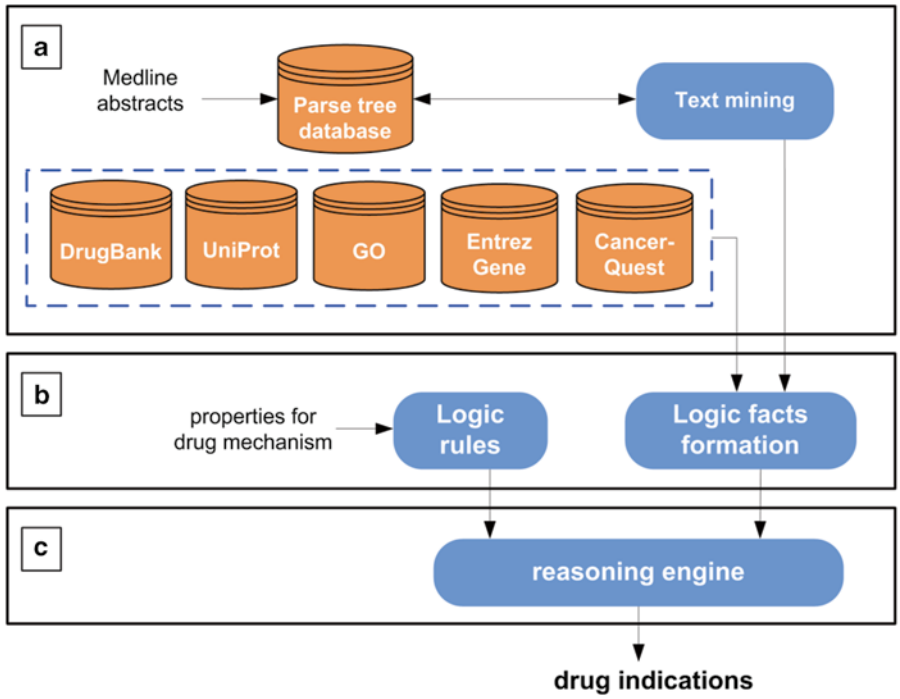

Fig. 1 An overview of the components involved in identifying alternative drug indications through text mining. These components include (**a**) the knowledge acquisition component, (**b**) the knowledge representation component, and (**c**) the reasoning component

3.1 Knowledge Representation

Basic drug mechanisms include the modulation, either activation or inhibition, of a protein target that is responsible for disease. These drug–target interactions then translate into clinical effects. For example, erlotinib is an epidermal growth factor receptor (EGFR) antagonist which alters the oncogenic EGFR signal transduction pathway. The key to identifying alterative drug indications is based on the principle that drug targets can be involved in diseases other than the original drug indication. Although compounds may interact with multiple targets, the primary target usually determines the first indication for development. Novel drug indications can be hypothesized through identifying alternative relations between primary targets and diseases as well as examining the secondary targets and their corresponding roles in disease.

Drug action representation involves initially identifying antagonists as triggering drug target inhibition and agonists as initiating drug target activation. With rich knowledge about cancer and its mechanisms, we applied our approach in identifying drugs that can be used as cancer treatments. A drug is identified as a treatment for cancer in one of the following scenarios:

- When the drug inhibits a protein that is known to be an oncogene.
- When the drug induces a protein that is known to be a tumor suppressor.

- When the drug inhibits a protein that is involved in a tumor-promoting biological process.
- When the drug induces a protein that is involved in a tumor-suppressing biological process.

Alternatively, a drug can also be identified as a cancer treatment when the drug activates protein function leading to an increase in tumor-suppressing protein expression or activates protein function leading to a decrease in tumor-promoting protein expression.

3.2 Relationship Extraction

While databases such as PharmGKB [17] and IntAct [18] are great resources for gene–disease relations and protein–protein interactions, they are limited in literature coverage due to the time-intensive process involved in manual curation. More importantly, it is commonly the case that interaction types are not captured in these databases. The information deficiency becomes an obstacle when the interactions from these databases are used in new knowledge discovery. As an example, let us assume that we know that protein *A* interacts with oncoprotein *B*. The interaction consequence (e.g., whether the function of *B* is ultimately activated or suppressed by *A*) is an important factor when the interaction is considered as a cancer drug treatment mechanism. To capture the interactions, we utilize text mining so that appropriate interactions can be identified efficiently from the literature.

Our text mining approach relies on grammatical structures and keywords to capture the directionality and the interaction types during gene–disease relation and protein–protein interaction extractions. The *parse tree query language (PTQL)* [19] is a suitable language that allows extraction patterns to be defined over keywords and grammatical structures. PTQL is designed for information extraction over a database of text known as the *parse tree database (PTDB)*. A parse tree is composed of a constituent tree and a linkage. A constituent tree is a syntactic sentence tree with the nodes represented by parts-of-speech tags and words in the sentence. A linkage represents the syntactic dependencies (or links) between word pairs in a sentence. The Stanford parser [20] is utilized to create parse sentence trees. BANNER [21] is used for gene name recognition from text, and the recognized gene names are then mapped to official gene symbols using GNAT [22]. The syntactic and semantic information is stored in our parse tree database, and extraction is performed by database queries. By storing the syntactic and semantic information, document collection reprocessing for every extraction is avoided. On-the-fly extraction is suitable for mining various interaction types as needed for identifying alternative drug indications.

A PTQL query is composed of four components: (1) tree patterns, (2) link conditions, (3) proximity conditions, and (4) return expression. The components in a PTQL query are separated

by the symbol ":". Here we describe the PTQL query syntax by the following query:

```
//S{ /NP{ //?[ Value='high'] => //?[ Value='levels'] =>
//?[ Tag='GENE'] (kw1)} => /VP{ //?[ Tag='DISE'] (kw2)}} :::
distinct kw1.value, kw2.value
```

The above tree query pattern specifies that within a noun phrase (denoted as `NP`), a gene name (denoted as variable `kw1`) has to be preceded by keywords "high" and "levels" through the operator =>. This gene name also needs to be followed by a verb phrase (denoted as `VP`), which contains a disease name mention (denoted as `kw2`). With the PTQL query, we obtained the relation that ADA overexpression is associated with acute lymphoblastic leukemia from the following sentence:

High levels of *adenosine deaminase (ADA)* activity have been associated with normal T cell differentiation and T cell disease, such as *acute lymphoblastic leukemia* (PMID: 6981287).

Readers can refer to [19] for more detailed information on PTQL and its implementation. By defining the keywords and extraction patterns over parse trees in the form of PTQL queries, it becomes possible to extract not only the interactions but also the directionality and interaction types. Specifically, the following interaction types are extracted:

1. Association between overexpressed or underexpressed genes and diseases.
2. Protein stimulation or inhibition by other proteins.

Sample interactions are listed in Table 1.

3.3 Logic Forms

In order to identify drug indications through automated reasoning, it is important to have proper drug mechanism knowledge representation. *Logic facts* are formed based on the knowledge acquired from the various sources as described in the previous subsection. In addition, *logic rules* are used to represent drug mechanism properties. We adopted a popular knowledge representation language called answer set programming (ASP) [23, 24] for logic fact and rule representation.

ASP is a declarative language that is useful for reasoning, including reasoning with incomplete information. An advantage for using a declarative language is that we define what the program should achieve and not how it should be achieved. Here we give a brief introduction to ASP syntax.

An *ASP rule* is in the form

$$l \leftarrow l_0, \ldots, l_m, \mathbf{not} l_{m+1}, \ldots, \mathbf{not} l_n$$

where l_is are literals and **not** represents *default negation*. The intuitive meaning of the above rule is that if it is known that literals l_0, …, l_m

Table 1
Sample extracted gene–disease relationships and protein–protein interactions with their support evidences

Evidences	Extracted relationships
The results of our study demonstrate that *AMACR* expression is *upregulated* in *gastric cancer* (PMID: 18787636)	<overexpressed AMACR, associated with, gastric cancer>
Therefore, *inactivation* of *Rb protein* by HPV 18 E7 protein may be associated with carcinogenesis of *small-cell carcinoma* (PMID: 14506638)	<underexpressed RB1, associated with, small cell carcinoma>
Moreover, *HER-2* expression was *stimulated* by *EGF* addition in young cells (PMID: 8028398)	<EGF, induces, ERBB2>
Inhibition of *PPARgamma* activity by *TNF-alpha* is involved in pathogenesis of insulin resistance (PMID: 18655773)	<TNF, inhibits, PPARG>

are to be true and if l_{m+1}, …, l_n are assumed to be false, then l must be true. A literal is defined as either an atom or an atom preceded by the symbol ¬ that indicates *classical negation*. If there is no literal l in the rule *head*, then the rule is referred to as a *constraint*. On the other hand, if there are no literals in the rule *body*, then the rule is referred to as a *fact*, and its representation fact short hand is simply the head literal itself. A set of ASP rules composes an answer set program, and an answer set program interpretation is called an answer set. Readers can refer to [25] for more details on ASP syntax and semantics.

Two basic logic fact types are used to represent the drug mechanisms: (1) concepts such as proteins and drugs and (2) interactions such as gene–disease relationships. The concept protein is represented in the *protein(Prot)* form, where *Prot* is a concept variable. For example, *protein(tp53)* indicates that tp53 is a protein concept instance. A complete concept and logic forms list is shown in Table 2. Interactions are represented with the predicate *interaction* for drug–target and protein–protein interactions and *relation* for gene–disease and gene-biological process relations. For instance, the logic form *relation(overexpressed(amacr), associated_with, gastric_cancer)* translates to overexpressed AMACR is associated with gastric cancer, and the logic form *interaction(egf, induces, erbb2)* represents that EGF induces ERBB2 activity. Table 3 shows a complete list of interaction types and their corresponding logic forms.

3.4 Automated Reasoning

With the knowledge denoted in logic fact form, drug mechanisms now need to be represented using ASP rules. The idea is to characterize and encode the mechanisms as pre- and post-interaction conditions, in which the precondition is represented in the body of

Table 2
Logic forms for the classes and entities involved in the drug mechanism domain

Facts	Logic forms	Examples
Prot is a protein, e.g., P53	*protein(Prot)*	*protein(tp53)*
Prot is an oncogene, e.g., EGFR	*oncogene(Prot)*	*oncogene(egfr)*
Prot is a tumor suppressor, e.g., P53	*suppressor(Prot)*	*suppressor(tp53)*
Dr is a drug, e.g., moclobemide	*drug(Dr)*	*drug(moclobemide)*
Dise is a disease, e.g., depression	*disease(Dise)*	*disease(depression)*
Bp is a cancer-promoting biological process, e.g., positive regulation of cell proliferation	*cancer_promoting_bioprocess(Bp)*	*cancer_promoting_bioprocess (pos_reg_cell_proliferation)*
Bp is a cancer-resisting biological process, e.g., positive regulation of apoptosis	*cancer_resisting_bioprocess(Bp)*	*cancer_resisting_bioprocess (pos_reg_apoptosis)*

Table 3
Logic forms for the interactions involved in the drug mechanism domain

Relations	Logic forms
Drug *Dr* induces the activity of protein *Prot*	*interaction(Dr, induces, Prot)*
Drug *Dr* inhibits the activity of protein *Prot*	*interaction(Dr, inhibits, Prot)*
Protein *Prot1* induces the activity of protein *Prot2*	*interaction(Prot1, induces, Prot2)*
Protein *Prot1* inhibits the activity of protein *Prot2*	*interaction(Prot1, inhibits, Prot2)*
Overexpressed protein *Prot* is associated with disease *Dise*	*relation(overexpressed(Prot), associated_with, Dise)*
Underexpressed protein *Prot* is associated with disease *Dise*	*relation(underexpressed(Prot), associated_with, Dise)*
Protein *Prot* plays a role in biological process *Bp*	*relation(Prot, is_associated, Bp)*

an ASP rule while the head represents the post-condition. Three rule sets are needed to perform inference for alterative drug indications: *initial triggers*, *inference rules*, and *constraints*. The initial triggers specify the criteria to initiate an inference. For drug mechanisms, the initial triggers correspond to drug target activation or inactivation by agonists or antagonists, respectively. The triggers are captured by the following rules:

- Initial trigger 1: Drug *Dr* activates drug target *Prot* function when *Dr* induces *Prot* expression:

  trigger(Dr, activates, Prot, 1) ← *interaction(Dr, induces, Prot)*, *protein(Prot)*, *drug(Dr)*.

- Initial trigger 2: Drug *Dr* inactivates drug target *Prot* function when *Dr* inhibits *Prot* expression:

  trigger(Dr, inactivates, Prot, 1) ← *interaction(Dr, inhibits, Prot), protein(Prot), drug(Dr).*

The following rules are used to represent other types of direct inference:

- Inference rule 1: Cancer is identified as an indication for drug *Dr* in step *S*+1 when tumor-suppressor *Prot* has been activated in the previous step *S*:

  trigger(Dr, treats, cancer, S+1) ← *trigger(Dr, activates, Prot, S), suppressor(Prot), drug(Dr), step(S).*

- Inference rule 2: Cancer is identified as an indication for drug *Dr* in step *S*+1 when oncogene *Prot* has been inhibited in the previous step *S*:

  trigger(Dr, treats, cancer, S+1) ← *trigger(Dr, inactivates, Prot, S), oncogene(Prot), drug(Dr), step(S).*

- Inference rule 3: Cancer is identified as an indication for drug *Dr* in step *S*+1 when protein *Prot*, which is involved in a cancer-promoting biological process *Bp*, has been inhibited in the previous step *S*:

  trigger(Dr, treats, cancer, S+1) ← *relation(Prot, is_associated, Bp), trigger(Dr, inactivates, Prot, S), protein(Prot), drug(Dr), cancer_promoting_bioprocess(Bp), step(S).*

- Inference rule 4: Cancer is identified as an indication for drug *Dr* in step *S*+1 when protein *Prot*, which is involved in a tumor-suppressing biological process *Bp*, has been activated in the previous step *S*:

  trigger(Dr, treats, cancer, S+1) ← *relation(Prot, is_associated, Bp), trigger(Dr, activates, Prot, S), protein(Prot), drug(Dr), cancer_resisting_bioprocess(Bp), step(S).*

- Inference rule 5: Cancer is identified as an indication for drug *Dr* in step *S*+1 when protein *Prot* has been inhibited in the previous step *S* and overexpressed *Prot* is known to be associated with cancer:

  trigger(Dr, treats, cancer, S+1) ← *trigger(Dr, inactivates, Prot, S), relation(overexpressed(Prot), associated_with, cancer), drug(Dr), protein(Prot), step(S).*

- Inference rule 6: Cancer is identified as an indication for drug *Dr* in step *S*+1 when protein *Prot* has been induced in the previous step *S* and underexpressed *Prot* is known to be associated with cancer:

  trigger(Dr, treats, cancer, S+1) ← *trigger(Dr, activates, Prot, S), relation(underexpressed(Prot), associated_with, cancer), drug(Dr), protein(Prot), step(S).*

In the above rules, the variable *S* is used to indicate a time stamp. Such time stamps represent interaction sequence, indicating that the different interactions that must occur prior to inferring a drug can be a cancer therapy. Such scenarios are considered *direct inferences* for cancer treatment. Furthermore, cancer therapies that are derived through drug-activated protein–protein interactions are considered *indirect inferences*, and they are represented with the ASP rules below:

- Inference rule 7: Drug *Dr* triggers protein *Prot2* functional activation in step *S*+1 when protein *Prot1* has been activated in the previous step *S* and activated *Prot1* increases *Prot2* expression:

  trigger(Dr, activates, Prot2, S+1) ← *trigger(Dr, activates, Prot1, S), interaction(Prot1, induces, Prot2), drug(Dr), protein(Prot1), protein(Prot2), step(S).*

- Inference rule 8: Drug *Dr* triggers protein *Prot2* functional inactivation in step *S*+1 when protein *Prot1* has been activated in the previous step *S* and activated *Prot1* decreases *Prot2* expression:

  trigger(Dr, inactivates, Prot2, S+1) ← *trigger(Dr, activates, Prot1, S), interaction(Prot1, inhibits, Prot2), drug(Dr), protein(Prot1), protein(Prot2), step(S).*

With the initial triggers and inference rules in place, constraints are used to define the valid inference criteria as follows:

- Constraint 1: An inference is valid only if goal becomes true, e.g., the series of steps must include the inference for a drug to be used as a cancer treatment, which is indicated by *trigger(Dr, treats, cancer, S)*:

  goal ← *trigger(Dr, treats, cancer, S), drug(Dr), step(S).*

  ← **not** *goal.*

- Constraint 2: No other interactions should follow *trigger(Dr, treats, cancer, S)* in a valid inference, ensuring that *trigger(Dr, treats, cancer, S)* is the last valid inference step:

  ← *trigger(Dr, activates, Prot, S), trigger(Dr, treats, cancer, S1), protein(Prot), drug(Dr), step(S), step(S1), S*≥ *S1.*

  ← *trigger(Dr, inactivates, Prot, S), trigger(Dr, treats, cancer, S1), protein(Prot), drug(Dr), step(S), step(S1), S*≥ *S1.*

To compute the answer sets that infer drug indications, an ASP solver called clingo [26] is utilized to compute direct and indirect inferences based on the rules and the acquired logic facts.

3.5 Dipyridamole as a Treatment for Cancer

Here we use the drug dipyridamole as an example to illustrate the direct inference of drug indications. Dipyridamole is prescribed to reduce blood clots through ADA inhibition [source: PubMed Health]. To find alternative indications for dipyridamole, we first acquire the necessary knowledge such as drug–target interactions and gene–disease relations. In this case, the following fact is acquired:

- Dipyridamole acts as an antagonist for ADA [source: DrugBank]:

  interaction(dipyridamole, inhibits, ada).

This interaction acts as the precondition for ADA functional inhibition through initial trigger 2, which results in *trigger(dipyridamole, inactivates, ada, 1)* being true. Mining biomedical abstracts reveals that ADA overexpression is associated with cancers like acute lymphoblastic leukemia.

- *High levels* of *ADA* activity have been *associated with* normal T cell differentiation and T cell disease, such as *acute lymphoblastic leukemia* [source: PMID: 6981287]:

  relation(overexpressed(ada), associated_with, cancer).

With the above interactions, inference rule 5 turns *trigger(dipyridamole, treats, cancer, 2)* to be true, indicating that dipyridamole is proposed as a potential treatment for cancer as ADA can be inhibited by dipyridamole and ADA overexpression is associated with cancer. This hypothesis together with the drug mechanisms is illustrated in Fig. 2.

3.6 Tazarotene as a Cancer Therapy

Here we use the drug tazarotene as an illustration for drug indication indirect inference. Tazarotene is approved for psoriasis and acne treatment. The facts below are acquired from different sources to identify an alternative indication for tazarotene.

- Tazarotene acts as an agonist for retinoic acid receptor alpha (RARA) [source: DrugBank]:

  interaction(tazarotene, induces, rara).

The above interaction results in *trigger(tazarotene, activates, RARA, 1)* to be true through initial trigger 1. By mining biomedical abstracts, it is discovered that RARA is known to inhibit EGFR oncogenic activity.

- These results suggest that RAR ligand-associated downregulation of EGFR activity reduces cell proliferation by reducing the magnitude and duration of EGF-dependent ERK1/2 activation [source: PMID: 11788593]:

  interaction(rara, inhibits, egfr).

- EGFR is a known oncogene [source: CancerQuest]:

  oncogene(efgr).

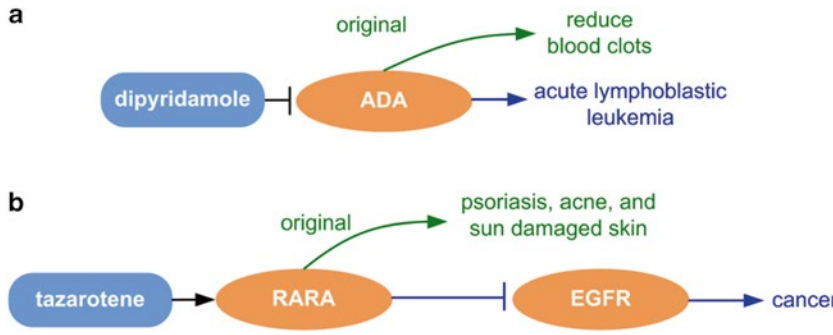

Fig. 2 A diagrammatic view of (**a**) direct and (**b**) indirect inferences for dipyridamole and tazarotene novel cancer indications

With the acquired facts and inference rule 8, *trigger(tazarotene, inactivates, EGFR, 2)* becomes true and subsequently turns *trigger(tazarotene, treats, cancer, 3)* to be true based on inference rule 2. The hypothesis generated through indirect inference indicates that the agonist tazarotene activates RARA which in turn inhibits EGFR, indicating the potential use of tazarotene as an oncology therapy. This hypothesis together with the drug mechanisms is illustrated in Fig. 2.

4 Conclusions

Drug repurposing plays an increasingly important role for pharmaceutical companies to minimize the time spent in the drug development process while maximizing previous investments. We described an approach that acquires knowledge from publicly available resources including Medline abstracts through text mining and generates alternative drug indication hypotheses through automated reasoning based on the acquired knowledge and drug mechanism logic representations. Using an evaluation set of 943 drugs obtained from DrugBank, 81 drugs are currently used for cancer treatments, while 289 drugs not having cancer as an original indication are currently being investigated as cancer therapies according to clinicaltrials.gov. Our method suggested 507 drugs that have the potential to be used for cancer treatments with a subset of 211 confirmed to be cancer related. Further analysis revealed that our approach was able to make 67 suggestions for cancer therapies among the 81 known cancer drugs (a recall of 82.7 %), and the remaining 144 suggestions are non-oncology drugs that are currently being tested in cancer clinical trials (a recall of 49.8 %). A more detailed result analysis can be found in [27].

It is important to note that there are a few important features that distinguish our approach from other literature-based approaches. These include (1) interaction-type extraction and utilization, (2) the

mining and application of directional interactions, and (3) the drug mechanism representation. Typical literature-based approaches that adopt the Swanson's ABC model usually produce large biological networks based on co-occurrence. Then researchers have to engage in the time-consuming process of using network visualization tools to sift through the networks and manually identify novel drug indications. The distinguishing features adopted in our approach reduce search space size so that it not only becomes computationally feasible, but, more importantly, the hypotheses generated reflect the drug mechanism of action as well as the key cancer mechanisms. These features lead to deriving potential alternative drug indications with scientific evidences explaining the mechanism behind the hypotheses.

The ability to identify alternative drug indications illustrates that combining text mining and automated reasoning is a powerful technique that enables knowledge inference in the biomedical domain.

References

1. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22:151–185. doi:10.1016/S0167-6296(02)00126-1
2. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3:673–683. doi:10.1038/nrd1468
3. Ghofrani HA, Osterloh IH, Grimminger F (2006) Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. Nat Rev Drug Discov 5:689–702. doi:10.1038/nrd2030
4. Dubus E, Ijjaali I, Barberan O, Petitet F (2009) Drug repositioning using in silico compound profiling. Future Med Chem 1:1723–1736. doi:10.4155/fmc.09.123
5. Yang L, Agarwal P (2011) Systematic drug repositioning based on clinical side-effects. PLoS One 6(12)
6. Duran-Frigola M, Aloy P (2012) Recycling side-effects into clinical markers for drug repositioning. Genome Med 4(1):3
7. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, Morgan AA, Sarwal MM, Pasricha PJ, Butte AJ (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 3:96ra76
8. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P (2013) Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 93(4):335–341
9. Deftereos SN, Andronis C, Friedla EJ, Persidis A, Persidis A (2011) Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdiscip Rev Syst Biol Med 3:323–334. doi:10.1002/wsbm.147
10. Swanson DR (1990) Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 78:29–37
11. Swanson DR (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30:7–18
12. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LTW, Aronson AR et al (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc 10:252–259. doi:10.1197/jamia.M1158
13. Yetisgen-Yildiz M, Pratt W (2006) Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 39:600–611. doi:10.1016/j.jbi.2005.11.010
14. Fu C, Jin G, Gao J, Zhu R, Ballesteros-Villagrana E, Wong ST (2013) DrugMap Central: an on-line query and visualization tool to facilitate drug repositioning studies. Bioinformatics 29(14):1834–1836. doi:10.1093/bioinformatics/btt279

15. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29
16. Knox C, Law V, Jewison T et al (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39(Database issue):D1035–D1041
17. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R et al (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1:167–170
18. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A et al (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Res 38:D525–D531. doi:10.1093/nar/gkp878
19. Tari L, Tu PH, Hakenberg J, Chen Y, Son TC et al (2010) Incremental information extraction using relational databases. IEEE Trans Knowledge Data Eng 24:86–99. doi:10.1109/TKDE.2010.214
20. Klein D, Manning CD (2003) Accurate unlexicalized parsing. Proceedings of the 41st Annual meeting on association for computational linguistics (ACL'03), Vol 1, pp 423–430. doi:10.3115/1075096.1075150
21. Leaman R, Gonzalez G (2008) BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. pp 652–663
22. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G (2008) Inter-species normalization of gene mentions with GNAT. Bioinformatics 24:i126–i132. doi:10.1093/bioinformatics/btn299
23. Gelfond M, Lifschitz V (1988) The stable model semantics for logic programming. In International symposium on logic programming, pp 1070–1080
24. Gelfond M, Lifschitz V (1991) Classical negation in logic programs and disjunctive databases. New Generation Computing 9:365–387
25. Baral C (2003) Knowledge representation, reasoning and declarative problem solving. Cambridge University Press, New York
26. Gebser M, Ostrowski M, Schaub T (2009) Constraint answer set solving. In Proceedings of the 25th International conference on logic programming (ICLP'09), Vol 5649, pp 235–249. doi:10.1007/978-3-642-02846-5
27. Tari L, Vo N, Liang S, Patel J, Baral C, Cai J (2012) Identifying novel drug indications through automated reasoning. PLoS One 7(7):e40946