Evidence based computational drug repositioning candidate screening pipeline design: Case Study

Qian Zhu[1]
[1]Department of Information Systems
University of Maryland, Baltimore County
Baltimore, Maryland, USA
qianzhu@umbc.edu

Hongfang Liu[2]
[2]Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota, USA
liu.hongfang@mayo.edu

Yuji Zhang[3]
[3]School of Medicine
University of Maryland
Baltimore, Mayland, USA
yuzhang@som.umaryland.edu

Jiabei Wang[4]
[4]School of Pharmacy
University of Maryland
Baltimore, Maryland, USA
jwang@rx.umaryland.edu

***Abstract*—Traditional drug development is time and cost consuming process, conversely, drug repositioning is an emerging approach to discover novel usages of existing drugs with a better risk-versus-reward trade-off. Computational technology is playing a key role in drug repositioning to screening the best drug repositioning candidates from a large candidate library. Recent efforts made for computer aided drug repositioning are mostly focusing on applying/developing data mining algorithms against wild type of large scale of biomedical data. In this paper, we introduce a novel computational pipeline designed for drug repositioning candidate screening based on existing phenotypical association (disease-disease association) discovery and pathway enrichment analysis by exploring systems biology data relevant to the interested phenotypical association specifically. To demonstrate usability and evaluate efficacy of this novel pipeline, we successfully conducted a case study by identifying potential drug repositioning candidates for Alzheimer's disease (AD) based on the studied phenotypical association between cancer and AD.**

Keywords—drug repositioning, phenotypical association, pathway enrichment analysis, systems biology

I. INTRODUCTION

By conservative estimates, it now takes about 15 years [1] and around $800 million to $1 billion to make a new drug to market[2], although drug development life cycle has significantly declined in recent decades owing to fast growth in drug research and development (R&D) such as chemical genomics technologies [3] [4] and chemical libraries[5] [6]. In another hand, an emerged novel strategy, discovering alternative usages for existing drugs, or drug repositioning, has been applied for decades[7]. While drug repositioning offers a better risk-versus-reward trade-off solution compared to traditional drug development, current successes in drug repositioning have primarily been the result of serendipity or clinical observations[8], such as the observed usefulness of sildenafil for erectile dysfunction and pulmonary arterial hypertension[9], as well as the new indications, including leprosy[10] and multiple myeloma[11] for thalidomide[12]. Systematic approaches by applying computational technologies have more capabilities to explore additional repositioning opportunities.

As the ability to measure molecules in high-throughput ways has improved over the past decade, it is logical that such data might be useful for enabling drug repositioning through computational methods. Many computational predictions for new indications have been borne out based on either drug or disease orientated strategies[13], and they are focusing on leveraging a large-scale of data and advanced informatics approaches to identify possible candidates for drug repositioning purpose. Such as, Andronis ed. al [14] have attempted to "integrate literature mining with other types of data arising from the use of these technologies as well as visualization tools assisting in the discovery of novel associations between existing drugs and new indications"; More other work by applying informatics approaches and machine learning prediction to detect novel usages of existing drugs from a large volume of chemical, biological data, genomic data, etc. has been published. [15-18] Although applying wild type of data gives more room for more possibilities to identify drug repositioning candidates, but it may not be the best way to seek repositioning candidates

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

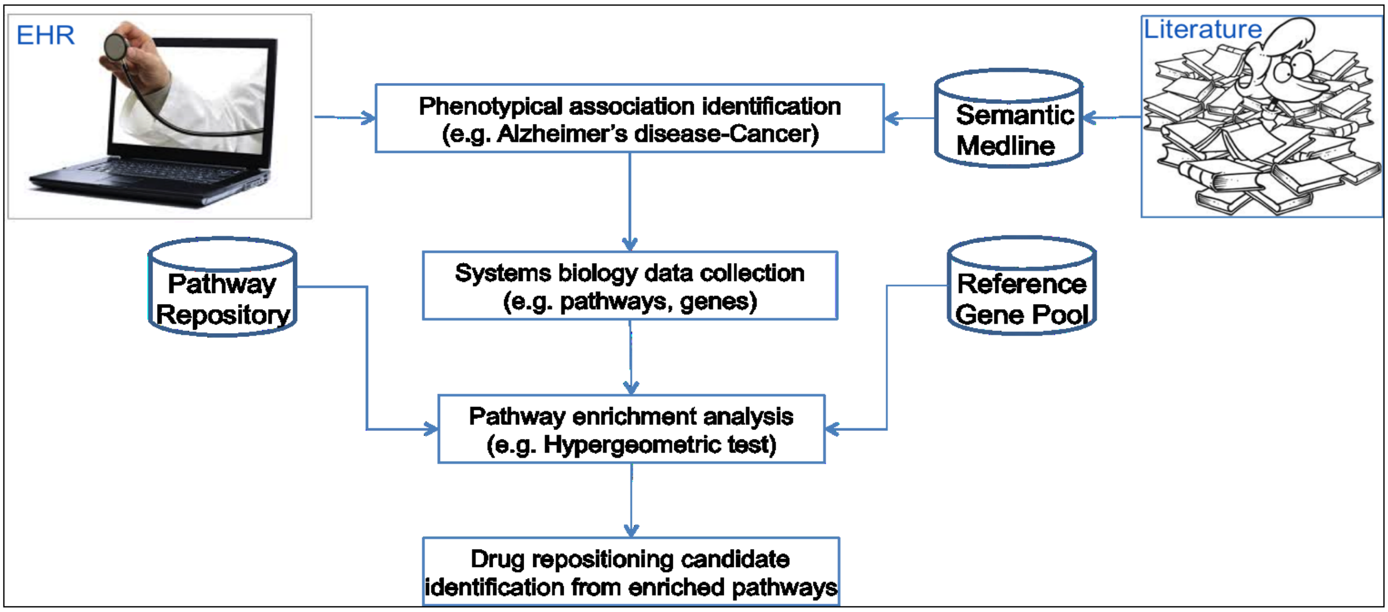
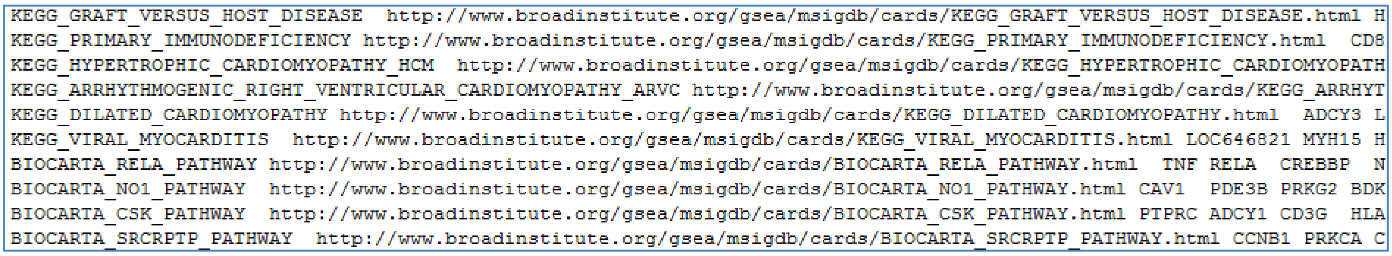

EHR
Literature
Phenotypical association identification
(e.g. Alzheimer's disease-Cancer)
Semantic Medline
Pathway Repository
Systems biology data collection
(e.g. pathways, genes)
Reference Gene Pool
Pathway enrichment analysis
(e.g. Hypergeometric test)
Drug repositioning candidate
identification from enriched pathways

Figure 1. Drug repositioning pipeline



```
KEGG_GRAFT_VERSUS_HOST_DISEASE  http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_GRAFT_VERSUS_HOST_DISEASE.html H
KEGG_PRIMARY_IMMUNODEFICIENCY http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_PRIMARY_IMMUNODEFICIENCY.html  CD8
KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM  http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_HYPERTROPHIC_CARDIOMYOPATH
KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_ARRHYT
KEGG_DILATED_CARDIOMYOPATHY http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_DILATED_CARDIOMYOPATHY.html  ADCY3 L
KEGG_VIRAL_MYOCARDITIS  http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_VIRAL_MYOCARDITIS.html LOC646821 MYH15 H
BIOCARTA_RELA_PATHWAY http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_RELA_PATHWAY.html  TNF RELA  CREBBP  N
BIOCARTA_NO1_PATHWAY  http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_NO1_PATHWAY.html CAV1  PDE3B PRKG2 BDK
BIOCARTA_CSK_PATHWAY  http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_CSK_PATHWAY.html PTPRC ADCY1 CD3G  HLA
BIOCARTA_SRCRPTP_PATHWAY  http://www.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_SRCRPTP_PATHWAY.html CCNB1 PRKCA C
```

Figure 2. Pathway examples

driven by specific interests, especially for finding drugs that can be repositioned for certain types of interested disease, such as cancer, depression. As we know underneath mechanism of drug repositioning is to manage associations between one drug and two types of disease, for example, determining whether one drug used for disease A can also be used for disease B. Thus starting with data that is relevant to one interested phenotypical association between disease A and disease B, will obviously provide more opportunities with higher successful repositioning rate. For instance, existing evidence is shown that cancer and Alzheimer's disease (AD) has association (more details can be found in Section 4 - case study), which allows us to mine data specifically related to cancer and AD, and consequently to find alternative drugs, either cancer drugs that were avoided in the case study presented in this study as cancer drugs with higher toxicity may not be suitable for older people or cancer related drugs for AD treatment. In this study, we introduce a phenotypical evidence based drug repositioning pipeline.

Pathways comprising genes and proteins may account for biological processes or diseases, which could be affected by a drug interacting with its pharmacological targets. Therefore, the pharmacological or clinical effects of a drug may be elucidated by analyzing the pathways enriched by drug targets with affinities or being affected by the studied drug candidates. There are some works being published recently by exploring pathway information for drug repositioning, Li et al.[19] developed a computational method for discovering new uses of existing drugs based on casual inference in a layered drug-target-pathway-gene-disease network. They simultaneously considered all possible causal chains connecting drugs to diseases via target- and gene-involved pathways. Pan et al. [20] investigated sixteen FDA-approved drugs for their mechanisms of action (MOAs) and clinical functions by pathway analysis based on retrieved drug targets interacting with or affected by the investigated drugs. They have illustrated some alternative therapeutic usages being found for these 16 drugs in their study. These published works started from interested drugs and attempted to find enriched pathways to these drugs, ultimately identify possible alternative therapeutic treatments for these drugs. In our work, we integrated a pathway enrichment analysis component into the drug repositioning pipeline. From those enriched pathways corresponding to the interested phenotypes, we will be able to identify drug repositioning candidates.

In this paper, we present our development and experiment for a novel computational drug repositioning pipeline. We begin with phenotypical association discovery, step-wise pathway enrichment analysis along with data preparation, to ultimately generate drug repositioning candidate library for the

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

interested phenotype. Details for each step with an example are described as below.

II. DRUG REPOSITIONING CANDIDATE SCREENING PIPELINE DESIGN

In this study, we introduced a generic drug repositioning screening pipeline designed for drug repositioning candidate screening based on established phenotypical associations. Specifically, this pipeline is consisting of several components, 1) phenotypical association identification via either literature scan, EHR data analysis, or other well-studied evidence; 2) systems biology data collection from known data resources, such as pharmacogenomics data, particularly for genes and pathways relevant to the phenotypical association identified from the first component; In parallel, human gene reference repository is built for further pathway enrichment analysis; 3) pathway enrichment analysis to identify possible drug repositioning candidates. Drug repositioning pipeline is shown in Figure 1 (shown in the last page). We will describe details in the following sections for each individual component.

A. Phenotypical evidence identification

There are multiple ways to identify possible phenotypical associations, either from literature, EHR data, or even from social media. In this paper, we will emphasize with literature and EHR, along with the case study by exploring literature data.

Literature provides a comprehensive published resource to identify possible phenotypical associations. Besides manually reviewing and automatically processing via Natural Language Processing from literature, Semantic Medline [21] is a centralized database of semantic predications from all PubMed citations. It includes associations among different nodes, such as drugs, genes, diseases and etc. The predicates express the associations identified from the literature between two nodes, for example, interacts_with, inhibit, stimulates, etc. Associations between two types of diseases can be identified and labeled with a PubMed identifier as reference.

Electronic Health Records (EHR) maintains a wide spectrum of patient information, including billing data, laboratory test results, medication records, clinical documentation and imaging results. It is likely that phenotypical associations can be identified from EHR via longitudinal patient data scan and analysis. For instance, Hua Xu et al. [22] has reported that metformin used to control blood sugar in patients with type 2 diabetes had better 5-year cancer survival rates compared to diabetic patients taking other diabetes medications for diabetic patients by linking a tumor registry to a large EHR database. While the authors have illustrated that metformin could be one drug repositioning candidate for cancer, such possible phenotypical association identified from EHR, diabetes and cancer can be applied into our designed pipeline, not only for finding further evidence to support the above finding, also identifying more possible candidates, besides metformin.

B. Recommended systems biology data resources

Pathway information including pathway names along with associated gene sets is recommended to be collected from Molecular Signatures Database (MSigDB) [23], which is a collection of annotated gene sets for use with GSEA software. Each pathway includes pathway name with embedded source name, a web link to MSigDB for more details about pathways and a list of genes involved. For instance, "BIOCARTA_RELA_PATHWAY" reflects to the pathway named "RELA" from BioCarta. More examples extracted from MSigDB are shown in Figure 2 (in the last page). Three major pathway resources listed below are included in MSigDB.

Kyoto Encyclopedia of Genes and Genomes (KEGG)[24] is a collection of manually drawn pathway maps expressing knowledge regarding to the molecular interaction and reaction networks, especially for Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases, Drug Development.

Reactome[25] is a manually curated and peer-reviewed pathway database. Pathway annotations are authored by biological experts, in collaboration with Reactome editorial staffs and cross-referenced to many bioinformatics databases. The core unit of the Reactome data model is the reaction. Entities (nucleic acids, proteins, complexes, vaccines, anti-cancer theraputics and small molecules) participating in reactions form a network of biological interactions and are grouped into pathways.

Biocarta[26] contains a large number of pathways in organisms. Each pathway comes with a detailed description in text format, which gives more information and additions to the graphical representation, and helps to understand the pathway better. The graphical representations of pathways also contain the chemical structure of the substance involved.

C. Gene-disease association identification

In order to collect genes relevant to two different diseases according to the phenotypical association identified from section 2.1, pharmacogenomics data and well-known predictive data are recommended as target resources, from where we can extract interested gene information. Some of recommended data resources are shown below. Ultimately, a centralized gene set, will be generated from these resources relevant to the identified phenotypical association specifically.

Pharmacogenomics Knowledge Base (PharmGKB)[27] contains genomic, phenotype and clinical information collected from pharmacogenomics (PGx) studies. It provides information regarding variant annotations, drug-centered pathway, pharmacogenomic summaries, clinical annotations, PGx-based drug-dosing guidelines, and drug labels with PGx information.

DrugBank [28] is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.

CellMiner™ [29] is a web application that facilitates systems biology through the retrieval and integration of the molecular and pharmacological data sets for the NCI-60 cell lines. CellMiner provides pattern comparisons for a given list of drugs and genes, and produces an output matrix that

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

includes correlated drugs and genes for each of drugs and genes from the query list along with corresponding calculated correlation values. By default, CellMiner returns significantly correlated entities (correlation value > 0.5) for a given query drug or gene.

D. Reference gene pool

In order to conduct enrichment analysis, a reference gene pool including all annotated human genes in HG 19 needs to be prepared.

E. Pathway enrichment analysis

Hypergeometric test is being applied for pathway enrichment analysis to identify possible drug repositioning candidates with significant correlations to phenotypical associations from enriched pathways. A five-step approach to perform hypergeometric test is proposed as below.

1) To ensure the reference gene pool (section 2.D) is a superset of the centralized gene set generated in section 2.C for pathway enrichment analysis, all genes from this centralized gene set should be mapped to the reference gene pool and all genes that were not included in the reference gene pool should be excluded. A gene set consisting of the remaining genes from the centralized gene set is called as CN_Gene.
2) For each pathway, a number of involved genes is counted and called as N_gp. Meanwhile, a number of genes from each pathway mapped to the genes from the CN_Gene is counted and called as N_gpm.
3) From the reference gene pool, we randomly select a subset of genes called as Sub_Gene with the same number as N_gpm for each pathway 1000 times. A number of genes from the Sub_Gene being mapped to CN_Gene is counted as N_gsm.
4) Number of times (N_t) that N_gsm is equal or greater than N_gpm, is calculated.
5) To prioritize and evaluate the significance for enriched pathways to AD and cancer, we calculated P value by N_t/1000.

Pathways can be ranked based on P_values. The pathways with smaller P_values, except for 0, are considered as enriched pathways for the phenotypical associations. For certain use cases, we can customize a cutoff P_value to customize the enriched pathway list.

F. Drug repositioning candidate screening

A list of drugs that are extracted from those enriched pathways consists of possible drug repositioning candidate library. There are two ways to extract drugs or chemicals from the pathways. First of all, KEGG and Reactome provide drug and chemical list that are involved in the pathways, and then it will be easy to extract those chemicals and drugs directly. Biocarta provides image and description for each pathway, and manual process will be required, for example, reviewing the image and description to identify chemicals/drugs from the pathways. For those pathways without chemical/drug information, the alternative means will be proposed to search for corresponding drugs to those genes available in the enriched pathways from the collected systems biological data.

Case study – drug repositioning candidates library generation for Alzheimer's disease (AD)

AD is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks. In most people with AD, symptoms first appear after age 60. Experts suggest that as many as 5.1 million Americans may have AD. However, there is no cure for AD currently. Drug and non-drug treatments may help with both cognitive and behavioral symptoms. Researchers are looking for new treatments to alter the course of the disease and improve the quality of life for people with AD. Currently there are only 4 FDA approved drugs, donepezil, galantamine, memantine, rivastigmine being used for AD. In this study, we aimed to identify possible alternative drugs may be drug repositioning candidates for AD by applying the drug repositioning pipeline introduced in this paper. We will go over all steps from phenotypical association discovery for AD, to drug repositioning candidate screening for AD in the following sections.

G. Pheotypical association identification for AD

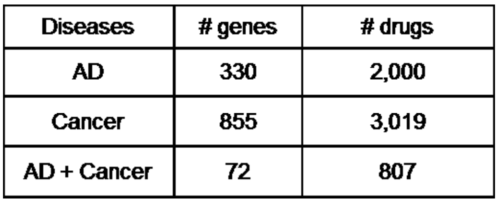
A study presented at the Alzheimer's Association International Conference® 2013 (AAIC® 2013) hold in Boston, including 3.5 million veterans reported that people who develop cancer appear to have a significantly reduced risk of developing AD, especially those who have had chemotherapy treatment[30]. Many other studies also supported such phenotypic association from different angles. White et al. [31], in a population-based longitudinal study recruiting 1,102 adults with a mean age of 79 years, showed that individuals older than 70 years of age with non-melanoma skin cancer had a significantly reduced risk of developing AD compared with those without non-melanoma skin cancer. Kamal et al. [32] revealed that NO(Nitric Oxide)-dependent abnormal mitochondrial activities and mitotic cell division are the important pathogenic factors in cancer and AD. Hedskog et al. [33] discovered that abnormal mitochondrial function was present in AD and cancer. Furthermore, some studies suggest chemotherapy may be beneficial for AD treatment. For instance, Cramer et al. [34] found that FDA approved anticancer drug bexarotene could be potentially used for AD treatment based on molecular pathway examination and analysis. Wang et al. [35] carried out a behavior screen in an AD fruit fly model and discovered that, the flies' memory improved after just two months of treatment of EGFR inhibitors (cancer drugs). In 2013, Araki [36] published a commentary on the potential of repositioning cancer drugs for the treatment of AD.

To get more proof for the studied phenotypical association between AD and cancer from a large biomedical literature repository, we searched for genes and drugs that relevant to AD and cancer from Semantic MEDLINE. The results are shown in Table 1. Clearly, there are large overlaps between genes associated with AD and cancer, and drugs associated with AD and cancer. However, the proportion of the overlaps is still quite small compared to the total number of genes and drugs. It is evident that the existence of this connection has

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

been supported by published studies, but the number of reports to investigate this connection is still fairly small, especially from informatics perspective. There are huge potentials to develop relevant studies to generate more evidence to support this phenotypical association from informatics point of view, which can guide experimental study design accordingly. The body of research outlined in this present paper leads to a hypothesis that connections between cancer and AD may show evidence to identify drug repositioning candidates for AD. This study demonstrates our desire and resources available to test this hypothesis.

TABLE Ⅰ. STATISTICAL SEARCH RESULTS FROM SEMANTIC MEDLINE



Diseases	# genes	# drugs
AD	330	2,000
Cancer	855	3,019
AD + Cancer	72	807

H. Systems biological data collection

Pathway set collection We downloaded the curated gene sets for pathways from KEGG, Reactome and BioCarta that are available at MSigDB by March 24, 2014. We collected pathway information consisting of pathway names along with associated gene sets from MSigDB. Total 186 KEGG pathways, 430 REACTOME pathways, and 217 BIOCARTA pathways have been retrieved from MSigDB and applied into this study subsequently.

Gene set collection We collected gene sets that are associated with cancer and AD. More specifically, four groups of gene sets have been retrieved, AD genes, cancer genes, AD related genes, and cancer related genes.

AD genes: we manually identified AD associated genes from four resources, PharmGKB [37], WikiPedia[38], NIH resource[39] and Literature[40]. Total 37 AD genes have been identified.

Cancer genes: Bushman lab [41] at university of Pennsylvania, has collected cancer related genes. All lists have been reconciled with current HGNC or NCBI gene IDs where outdated synonyms were used. We downloaded this comprehensive list of cancer related genes called allOnco for this study. 2,129 cancer genes have been extracted from allOnco table.

AD related genes: We explored CellMiner to identify genes with similar gene expression profile to AD genes. CellMiner produces an output matrix including correlated genes with correlation value. We extracted all genes with correlation value greater than 0.5 for each AD genes. Out of total 22,647 genes extracted from CellMiner, 3,739 genes significantly correlated (correlation value > 0.5) to 34 AD associated Genes were extracted from CellMiner output matrix. It notes that there are no correlated genes found for three AD genes, NME8, SCL24A4, and TERM2 from CellMiner. It is worthy to note that there are no correlated results generated for four AD drugs from CellMiner.

Cancer related genes: Total 128 unique cancer drugs with synonymies have been extracted from Medilexicon [42] by Nov. 16, 2013. To identify genes associated with those cancer drugs, we explored PharmGKB by May 8, 2013, which provides high quality of associations among drug, gene, disease, SNP and haplotype. We programmatically searched relevant genes for those 128 cancer drugs by using string matching with drug names and synonymies based on the drug-gene pairs available at PharmGKB relationship file. In parallel, we downloaded “External Links” and “Drug Target Identifiers” files from DrugBank by April 22, 2014 to identify genes that are associated with cancer drugs from DrugBank. We searched for cancer drug related genes via two steps from DrugBank, converting drug names to DrugBank identifiers based on “External Links” file and identifying relevant genes from “drug target identifiers” file by parsing the DrugBank identifiers. For the unmapped drugs with PharmGKB and DrugBank, we manually searched for relevant genes from literatures. In summary, we programmatically extracted 249 unique genes for 58 unique cancer drugs from PharmGKB and 134 unique genes for 96 unique cancer drugs from DrugBank. Combining the searching results from these two resources, total 346 unique genes have been found for 101 unique cancer drugs. Aggregating searching result based on manually literature review, total 412 unique genes for 115 cancer drugs have been applied in this study.

We generated a centralized gene set by combing the above four gene sets.

Reference gene pool generation 25,237 human genes were collected to consist of a human gene pool.

Table 2 summarizes number of genes from different gene collections listed as above have been applied in this study.

TABLE Ⅱ. NUMBERS OF GENES BEING APPLIED IN THIS CASE STUDY

Gene collections	Total # genes	# genes covered by the reference gene pool
AD genes	37	35
Cancer genes	2,129	1,975
AD related genes	3,739	3,045
Cancer related genes	412	241
Human gene pool	25,237	

I. Pathway enrichment analysis

To conduct enrichment analysis, we followed the step introduced in section 2.E. First of all, we removed genes from the centralized gene set that were not included in the reference gene pool and mapped them to gene sets for each pathway identified in section 3.B (pathway set collection). Finally, a centralized gene set has been downsized to 4,688 unique genes that relevant to AD and cancer.

The centralized gene set and the human gene reference pool have been applied for pathway enrichment analysis to identify significant correlated pathways to AD and cancer by following the steps 2-5 described in section 2.E. Each pathway has been assigned a p value indicating the significance to AD and cancer. By excluding those pathways with p-value equaling to 0 and greater than 0.05, total 193 pathways were selected for this case study. From there, drug repositioning candidate library for

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

AD can be generated to include drugs involved or associated with those pathways.

J. Evaluation

To evaluate and test our hypothesis – possible drug repositioning candidates can be screened out from the enriched pathways to AD, we selected top 24 enriched pathways with p-Value = 0.001, which include 3 pathways from KEGG, 10 from Biocarta, and another 11 from Reactome. The list of those 24 enriched pathways is shown in Table 3. From these pathways, we manually extracted available drugs/chemicals. It is worthy to note that not all pathways include chemicals/drugs. There are total 110 chemicals/drugs. KEGG and Reactome provide drug and chemical list that are involved in the pathways, and then we extracted those chemicals and drugs directly, and we manually reviewed the image and description of pathways from Biocarta to identify chemicals/drugs.

TABLE III. THE LIST OF TOP 24 ENRICHED PATHWAYS

KEGG_COMPLEMENT_AND_COAGULATION_CASCADES
KEGG_PRION_DISEASES
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS
BIOCARTA_NO1_PATHWAY
BIOCARTA_ASBCELL_PATHWAY
BIOCARTA_CD40_PATHWAY
BIOCARTA_GCR_PATHWAY
BIOCARTA_SKP2E2F_PATHWAY
BIOCARTA_IL5_PATHWAY
BIOCARTA_IL10_PATHWAY
BIOCARTA_IL12_PATHWAY
BIOCARTA_PTDINS_PATHWAY
BIOCARTA_BARR_MAPK_PATHWAY
REACTOME_G2_M_TRANSITION
REACTOME_HDL_MEDIATED_LIPID_TRANSPORT
REACTOME_NEURORANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL
REACTOME_NOTCH_HLH_TRANSCRIPTION_PATHWAY
REACTOME_NUCLEOTIDE_EXCISION_REPAIR
REACTOME_PLATELET_ADHESION_TO_EXPOSED_COLLAGEN
REACTOME_PLC_BETA_MEDIATED_EVENTS
REACTOME_SHC_RELATED_EVENTS
REACTOME_SIGNALING_BY_NOTCH
REACTOME_CREB_PHOPHORYLATION_THROUGH_THE_ACTIVATION_OF_RAS
REACTOME_POST_NMDA_RECEPTOR_ACTIVATION_EVENTS

Case 1. Biliverdin and Bilirubin are two chemicals inlcuded in "10 Anti-inflammatory Signaling Pathway" from Biocarta. Biliverdin and Bilirubin are green tetrapyrrolic bile pigments that naturally possess significant anti-mutagenic and antioxidant properties and therefore fulfill a useful physiological function[43]. Biliverdin and bilirubin have been shown to be potent scavengers of peroxyl radicals.[43, 44] They have also been shown to inhibit the effects of polycyclic aromatic hydrocarbons, heterocyclic amines, and oxidants. Studies have reported that people with higher concentration levels of bilirubin and biliverdin in their bodies have a lower frequency of cancer and cardiovascular disease[44].

In another hand, oxidative damage has been shown to be a factor in Alzheimer's disease (AD), and some studies have suggested that supplemental anti-oxidants can decrease the risk of AD or slow its progression. There are many candidate antioxidants, including combinations, which could be neuroprotective in established AD or could have efficacy in the prevention of AD. Biliverdin and Bilirubin could be the ones as candidate antioxidants for AD treatment. Barone et al. has illustrated that "Biliverdin reductase-A (BVR-A) is a pleiotropic enzyme and plays pivotal role in the antioxidant defense against free radicals as well as in cell homeostasis. Together with heme oxygenase, BVR-A forms a powerful system involved in the cell stress response during neurodegenerative disorders including Alzheimer's disease (AD), whereas due to the serine/threonine/tyrosine kinase activity the enzyme regulates glucose metabolism and cell proliferation"[45]. One clinical trial "Anti-Oxidant Treatment of Alzheimer's Disease" has been done to examine the safety and effectiveness of two anti-oxidant treatment regimens in patients with mild to moderate Alzheimer's disease[46].

Case 2. Dabigatran is one of drugs included in "COMPLEMENT AND COAGULATION CASCADES" pathway from KEGG. Dabigatran is an oral anticoagulant drug that acts as a direct thrombin (factor IIa) inhibitor[47]. Dabigatran can be used for the prevention of stroke in patients with atrial fibrillation. The drug was developed as an alternative to warfarin [47]. In addition, scientists recently found that dabigatran as direct thrombin inhibitors, "might be efficient in the treatment of patients with AD because of their high selectivity for thrombin's activity inhibition while having a safer side effects profile than heparin."[48] Ramasamy et al. [49] has also investigated the effectiveness of dabigatran in treating AD, Figure 3 shows the rationale they proposed in their paper, which presented the pathway how dabigatran can contribute to the development of AD. Other studies have made the same statements, such as "Dabigatran reduces expression of HIF-1α, thrombin, IL-6, MCP-1, and MMP2 in the brains of AD transgenic mice. Generation of ROS in AD mice and hypoxic endothelial cell cultures is inhibited by dabigatran"[50].

Among 110 chemicals/drugs, we evaluated 3 above chemicals/drugs as potential drug repositioning candidates for AD treatment based on existing studies and clinical trials. However, other chemicals/drugs can be served as new drug repositioning candidates that are worthy to be investigated further for repositioning purpose.

III. DISCUSSION

This study was aiming to build phenotypical evidence based computational drug repositioning candidate screening pipeline, along with a case study to demonstrate the capability of this pipeline to screen possible drug repositioning candidates for AD treatment. In the case study, given the fact of phenotypical association between AD and cancer, we successfully identified drug candidates from pathways that are significantly correlated to AD and cancer by performing pathway enrichment analysis. This approach dramatically decreased the traditional searching space and increased the success rate to identify drug repositioning candidates for AD. A drug repositioning candidate library has been generated by the presented approach, followed by the evaluation - manual evidence identification.

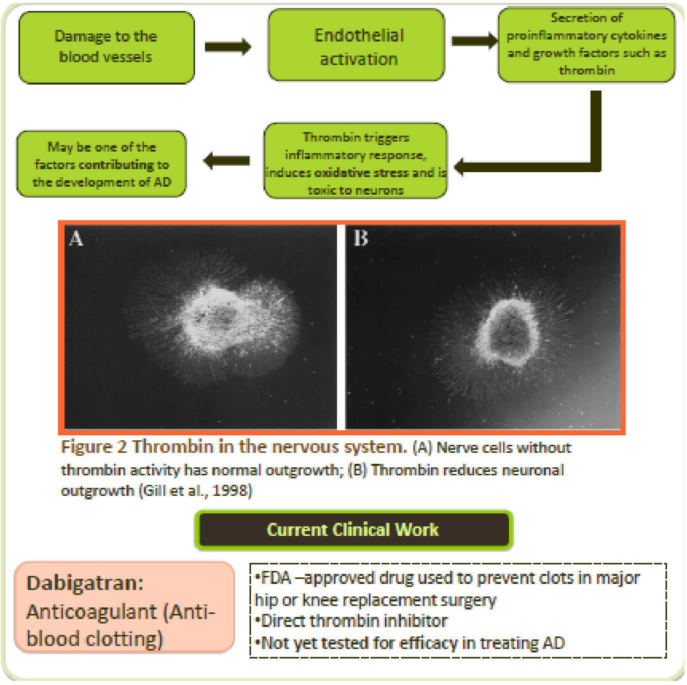

Damage to the blood vessels
Endothelial activation
Secretion of proinflammatory cytokines and growth factors such as thrombin
Thrombin triggers inflammatory response, induces oxidative stress and is toxic to neurons
May be one of the factors contributing to the development of AD
A
B
Figure 2 Thrombin in the nervous system. (A) Nerve cells without thrombin activity has normal outgrowth; (B) Thrombin reduces neuronal outgrowth (Gill et al., 1998)
Current Clinical Work
Dabigatran: Anticoagulant (Anti-blood clotting)
•FDA –approved drug used to prevent clots in major hip or knee replacement surgery
•Direct thrombin inhibitor
•Not yet tested for efficacy in treating AD

Figure 3. The rationale supporting dabigatran's contribution to AD development (Adopted from [49])

To validate our approach with different reference gene pool, we performed the same approach by using a different reference gene pool consisting of genes extracted from all selected pathways, and we have received similar results. As genes from pathways are more significantly relevant to AD and cancer, the number of enriched pathways is much larger comparing to use human genes as a reference gene pool. To avoid such unexpected increase of the number of enriched pathways, we decided to apply human genes as reference gene pool.

As following, we will discuss the benefits gained and consequent findings sought from this study, as well as limitations we observed from this study and subsequent future work.

Benefits gained

The current scenario of computational drug repositioning is based on a large scale of wild type of data to find possible drug repositioning candidates. Conversely, our strategy was driven by the existing phenotypical associations that provide solutions to find drug repositioning candidates for specific disease. For example, investigation conducted upon the data retrieval driven by the existing phenotypical association between AD and cancer demonstrates its capability of identifying possible drug repositioning candidates for AD specifically. Meanwhile, this approach not only increases the success rate for drug repositioning candidate discovery as it is supported by the existing studied phenotypical associations, but also it dramatically decreases size of the drug repositioning candidate screening pool as only drugs/chemical compounds included in the enriched pathways that are associated with interested diseases will be selected.

Drug repositioning candidates can be identified by applying our computational pipeline introduced in this paper via different ways, 1) we can directly look for candidates based on phenotypical associations. For example, searching for possible candidates repositioned for AD treatment against cancer drug pool. However, if we consider the toxic nature of the cancer drugs that may be not the best choice for elderly who has been diagnosed with AD, then we are able to switch to an alternative way, 2) we can look into the common pathways relevant to the interested phenotypical association, such as AD and cancer. Then we can find possible drug candidates from those pathways. Moreover, 3) it will be very easy to integrate other relevant resources to the information identified for the phenotypical association specifically into a network and conduct network analysis for drug repositioning candidate discovery.

Drug repositioning candidates identified from our pipeline will not only include ones have already been studied, which serves as evidence to demonstrate the efficacy and performance of our pipeline, but also include more novel candidates that have not been investigated before. Those novel findings will be the main contribution to the drug repositioning field as that may provide more new hints leading new discovery.

Limitations observed.

While we successfully demonstrated the promising findings and performance of this study, we observed some limitations of this study, and proposed relevant future work plan.

1) In the case study, our focus was identifying significant gene sets to AD and cancer and conducting pathway enrichment analysis to seek enriched pathways. We manually browsed the content of the enriched pathways and extracted annotated drug and chemical concepts, from which drug repositioning candidates for AD can be identified. For demonstration purpose, manual process provides high quality of identification results, but also it offers more guidance for future candidate identification from huge volume of data, for instance, more pathways from additional resources, like PharmGKB, wikiPathway and more reference gene sets extracted from pathways. At the time of increasing number of enriched pathways, systematical drug repositioning candidate identification process will be established, including automated drug/chemical extraction from pathways by applying Nature Language Processing (NLP) and prioritizing candidates by leveraging evaluation (discuss more in the next paragraph) and evidence found from the literature.

2) In this study, we manually identified drug repositioning candidates from the enriched pathways and found relevant evidence by reviewing literature. However, we were not employing further automated evaluation process to validate the possibility of repositioning that is beyond the scope of this

study, and is the future plan to design a computational application for drug repositioning by applying this presented approach in an automatic way. More specifically, EHRs maintain huge volume of patient medical information, such as current/past medication and diagnosis information. From there, we can conduct longitudinally retrospective study to find out the possibility of the identified drug candidates being used for real patients. Meanwhile, drug candidates identified for AD treatment can be further evaluated for druggability, blood brain barrier (BBB) penetration by applying several 'rules of thumb' that have emerged from studies [51-53] to give simple guidance concerning the molecular properties that favor brain permeation.

IV. Conclusion

This presented study has introduced a novel approach to identify possible drug candidates via phenotypical association discovery and pathway enrichment analysis. The presented case study has successfully demonstrated the capability of our approach being used for identifying drug repositioning candidates for AD. Evidence based phenotypical associations increase the success rate of searching for possible drug repositioning candidates by decreasing candidate searching space that is only associated with the interested phenotypical association. Integrating system biological information lowers the risk for repositioning existing chemicals or drugs as the candidates are extracted from the pathways. It is worthy to highlight that this approach can be extended to other interested disease areas driven by other existing phenotypic associations.

References

[1] DiMasi, J.A., New drug development in the United States from 1963 to 1999. Clinical pharmacology and therapeutics, 2001. 69(5): p. 286-296.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, The price of innovation: new estimates of drug development costs. Journal of health economics, 2003. 22(2): p. 151-185.

[3] Kim, D.-H. and T. Sim, Chemical kinomics: a powerful strategy for target deconvolution. BMB Rep, 2010. 43(11): p. 711-719.

[4] Roemer, T., et al., Bugs, drugs and chemical genomics. Nature chemical biology, 2012. 8(1): p. 46-56.

[5] Wang, Y., et al., PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic acids research, 2009. 37(suppl 2): p. W623-W633.

[6] Gaulton, A., et al., ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research, 2012. 40(D1): p. D1100-D1107.

[7] Ashburn, T.T. and K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov, 2004. 3(8): p. 673-683.

[8] Liu, Z., et al., < i> In silico</i> drug repositioning–what we need to know. Drug discovery today, 2013. 18(3): p. 110-115.

[9] Galiè, N., et al., Sildenafil citrate therapy for pulmonary arterial hypertension. New England Journal of Medicine, 2005. 353(20): p. 2148-2157.

[10] Silverman, W.A., The schizophrenic career of a "monster drug". Pediatrics, 2002. 110(2): p. 404-406.

[11] Singhal, S., et al., Antitumor activity of thalidomide in refractory multiple myeloma. New England Journal of Medicine, 1999. 341(21): p. 1565-1571.

[12] McBride, W.G., Thalidomide and congenital abnormalities. The Lancet, 1961. 278(7216): p. 1358.

[13] Dudley, J.T., T. Deshpande, and A.J. Butte, Exploiting drug–disease relationships for computational drug repositioning. Briefings in bioinformatics, 2011: p. bbr013.

[14] Andronis, C., et al., Literature mining, ontologies and information visualization for drug repurposing. Briefings in Bioinformatics, 2011. 12(4): p. 357-368.

[15] Moriaud, F., et al., Identify drug repurposing candidates by mining the Protein Data Bank. Briefings in Bioinformatics, 2011. 12(4): p. 336-340.

[16] Schneider, P., Y. Tanrikulu, and G. Schneider, Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. Current medicinal chemistry, 2009. 16(3): p. 258-266.

[17] Loging, W., et al., Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing. Drug Discovery Today: Therapeutic Strategies, 2012. 8(3): p. 109-116.

[18] Lussier, Y.A. and J.L. Chen, The Emergence of Genome-Based Drug Repositioning. Science Translational Medicine, 2011. 3(96): p. 96ps35.

[19] Li, J. and Z. Lu, Pathway-based drug repositioning using causal inference. BMC Bioinformatics, 2013. 14(Suppl 16): p. S3.

[20] Pan, Y., et al., Pathway Analysis for Drug Repositioning Based on Public Database Mining. Journal of Chemical Information and Modeling, 2014. 54(2): p. 407-418.

[21] Rindflesch, T.C., et al., Semantic MEDLINE: An advanced information management application for biomedicine. Information Services and Use, 2011. 31(1): p. 15-21.

[22] Xu, H., et al., Electronic health record data suggests metformin improves cancer survival: A new model for drug repurposing studies. AMIA 2012 Annual Symposium.

[23] Liberzon, A., et al., Molecular signatures database (MSigDB) 3.0. Bioinformatics, 2011. 27(12): p. 1739-1740.

[24] Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 2000. 28(1): p. 27-30.

[25] Joshi-Tope, G., et al., Reactome: a knowledgebase of biological pathways. Nucleic acids research, 2005. 33(suppl 1): p. D428-D432.

[26] Nishimura, D., BioCarta. Biotech Software & Internet Report: The Computer Software Journal for Scient, 2001. 2(3): p. 117-120.

[27] Hewett, M., et al., PharmGKB: the pharmacogenetics knowledge base. Nucleic acids research, 2002. 30(1): p. 163-165.

[28] Wishart, D.S., et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research, 2008. 36(suppl 1): p. D901-D906.

[29] Reinhold, W.C., et al., CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer research, 2012. 72(14): p. 3499-3511.D

[30] Cancer And Chemotherapy Linked With Decreased Risk Of Alzheimer's Disease In Veterans. [cited 2014 January 28]; Available from: http://www.prnewswire.com/news-releases/cancer-and-chemotherapy-linked-with-decreased-risk-of-alzheimers-disease-in-veterans-215502921.html.

[31] White, R.S., et al., Nonmelanoma skin cancer is associated with reduced Alzheimer disease risk. Neurology, 2013. 80(21): p. 1966-1972.

[32] Kamal, M.A., Link between Cancer and Alzheimer Disease via Oxidative Stress Induced by Nitric Oxide-Dependent Mitochondrial DNA Overproliferation and Deletion. Oxidative medicine and cellular longevity, 2013. 2013.

[33] Hedskog, L., S. Zhang, and M. Ankarcrona, Strategic Role for Mitochondria in Alzheimer's Disease and Cancer. Antioxidants & Redox Signaling, 2012. 16(12): p. 1476-1491.

[34] Cramer, P.E., et al., ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models. Science, 2012. 335(6075): p. 1503-1506.

[35] Wang, L., et al., Epidermal growth factor receptor is a preferred target for treating Amyloid-β–induced memory loss. Proceedings of the National Academy of Sciences, 2012.

[36] Araki, W., Potential repurposing of oncology drugs for the treatment of Alzheimer's disease. BMC Medicine, 2013. 11(1): p. 1-3.

[37] Related Genes for Alzheimer Disease. [cited 2014 Jan, 2]; Available from: https: Figure 2. Pathway examples

[38] Alzheimer's disease. [cited 2014 Jan,2]; Available from: http://en.wikipedia.org/wiki/Alzheimer%27s_disease#Genetics.

[39] Alzheimer disease. [cited 2014 Jan, 2]; Available from: http://ghr.nlm.nih.gov/condition/alzheimer-disease.

[40] Lambert, J.-C., et al., Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature genetics, 2013.

[41] Cancer gene list. [cited 2014 April 21]; Available from: http://www.bushmanlab.org/links/genelists.

[42] Cancer drugs & oncology drugs. [cited 2014 April 21]; Available from: http://www.medilexicon.com/drugs-list/cancer.php.

[43] Orui, T., et al., Transient Relief of Asthma Symptoms during Jaundice: A Possible Beneficial Role of Bilirubin. The Tohoku Journal of Experimental Medicine, 2003. 199(3): p. 193-196.

[44] Bulmer, A.C., et al., The anti-mutagenic properties of bile pigments. Mutation Research/Reviews in Mutation Research, 2008. 658(1–2): p. 28-41.

[45] Barone, E., et al., Oxidative and nitrosative modifications of biliverdin reductase-A in the brain of subjects with Alzheimer's disease and amnestic mild cognitive impairment. Journal of Alzheimer's Disease, 2011. 25(4): p. 623-633.

[46] Clinical trial for "Anti-Oxidant Treatment of Alzheimer's Disease". [cited 2014 June 23th]; Available from: https://clinicaltrials.gov/ct2/show/NCT00117403?term=antioxidants+AND+%22Alzheimer%27s+disease%22&rank=1.

[47] Dabigatran. [cited 2014 June 17th]; Available from: http://en.wikipedia.org/wiki/Dabigatran.

[48] Rami, B.K., Direct Thrombin Inhibitors' Potential Efficacy in Alzheimer's Disease. American Journal of Alzheimer's Disease and Other Dementias, 2012. 27(8): p. 564-567.

[49] Ramasamy, H.K. and T. Cheng, The Effectiveness of Dabigatran in Treating Alzheimer's Disease.

[50] Tripathy, D., et al., Thrombin, a mediator of cerebrovascular inflammation in AD and hypoxia. Frontiers in aging neuroscience, 2013. 5.

[51] Norinder, U. and M. Haeberlein, Computational approaches to the prediction of the blood–brain distribution. Advanced drug delivery reviews, 2002. 54(3): p. 291-313.

[52] Clark, D.E., In silico prediction of blood–brain barrier permeation. Drug Discovery Today, 2003. 8(20): p. 927-933.

[53] Pajouhesh, H. and G.R. Lenz, Medicinal chemical properties of successful central nervous system drugs. NeuroRx, 2005. 2(4): p. 541-553.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE