약학박사학위논문

신약재창출을 위한 계산 방법

- Computational Methods for Drug repositioning -

2013 년 2 월

서울대학교 약학대학원

약학과 의약생명과학전공

이 지 현

Abstract

The process of drug discovery and development is time-consuming and costly, and the probability of success is low. Therefore, there is rising interest in repositioning existing drugs for new medical indications. When successful, this process reduces the risk of failure and costs associated with de novo drug development. However, in many cases, new indications of existing drugs have been found serendipitously. Thus there is a clear need for establishment of rational methods for drug repositioning.

In this study, I have established a database we call “PharmDB” which integrates data associated with disease indications, drug development, and associated proteins, and known interactions extracted from various established databases. And the Shared Neighborhood Scoring (SNS) algorithm was designed to explore the inferred linkages of known drugs to diseases of interest.

I also developed a combinatorial drug discovery system called “CDA (Combinatorial Drug Assembler)”. CDA performs gene expression pattern matching of signaling pathway components to discover synergistic combinatorial drug pairs. Despite successes of several drugs targeting highly specific single disease-associated gene products, many of other drugs show inefficient effect or severe side effects. To

overcome these limitations of single protein targeting drug, CDA was developed. CDA provides a new way for rational combinatorial drug discovery.

The data in PharmDB is open access and can be easily explored with phExplorer and accessed via BioMart web service (http://www.i-pharm.org/, http://biomart.i-pharm.org/). CDA is freely available at http://cda.i-pharm.org.

Keywords: Drug repositioning, tripartite network, Shared Neighborhood Scoring (SNS) algorithm, systems biology, combinatorial drug, signaling pathway, gene expression

Student number: 2008-31001

Contents

Chapter I

Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug

Chapter II

CDA: Combinatorial drug discovery using transcriptional response modules

List of figures and tables

Chapter I

Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug

Chapter II

CDA: Combinatorial drug discovery using transcriptional response modules

Chapter I

Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug

Running title: Inferred relationships by considering shared neighborhood

Keywords: Drug repositioning, tripartite network, Shared Neighborhood Scoring (SNS) algorithm, systems biology

Abbreviations list

HTS: Highthrouput Screening

SNS: Shared Neighborhood Scoring

SN score: Shared Neighborhood score

MINT: the Molecular INTeraction database

DIP: the Database of Interacting Proteins

CTD: The Comparative Toxicogenomics Database

TTD: Therapeutic Target Database

PharmGKB: The Pharmacogenomics Knowledge Base

OMIM: Online Mendelian Inheritance in Man

GAD: Genetic Association Database

ROC: Receiver Operating Characteristic

AUC: Area Under Curve

AZA: Acetazolamide

CA: Carbonic anhydrase

SCC: Squamous Cell Carcinoma

TBZT: Thia-benzthiazide

KS test: Kolmogorov-Smirnov test

Abstract

Drug repositioning, also known as drug repurposing, is the process of discovery of exiting drug for new medical indications. Drug repositioning has been growing in importance since the drug development is a long, costly, and high-risk business. In this study, I have established a database called “PharmDB” which integrates data associated with disease indications, drug development, and associated proteins, and known interactions extracted from various established databases. To explore linkages of known drugs to diseases of interest from within PharmDB, the Shared Neighborhood Scoring (SNS) algorithm was designed. And to facilitate exploration of tripartite (Drug-Protein-Disease) network, I developed a graphical data visualization software program called phExplorer, which allows us to browse PharmDB data in an interactive and dynamic manner. I validated this knowledge-based tool kit, by identifying a potential application of a hypertension drug, benzthiazide (TBZT), to induce lung cancer cell death.

By combining PharmDB, an integrated tripartite database, with Shared Neighborhood Scoring (SNS) algorithm, I developed a knowledge platform to rationally identify new indications for known FDA approved drugs, which can be customized to

specific projects using manual curation. The data in PharmDB is open access and can be easily explored with phExplorer and accessed via BioMart web service (http://www.i-pharm.org/, http://biomart.i-pharm.org/).

Introduction

Modern drug discovery is time-consuming and expensive, involving coordinated multi-disciplinary research in multiple stages, each requiring intensive and specialized resources [1]. Although rapid advancement of "omics" approaches, computational systems biology and accumulation of digital data resources have provided a vast array of significant information in life science [2], data relevant to drug discovery are not easily identified and recruited for application to pharmaceutical research [3]. Despite the technological advances in drug discovery such as HTS, the approval of new drugs has remained stagnant in the past decade, resulting in an overall decline in the productivity of the pharmaceutical industry.

In efforts to save development time and minimize the risk of failure during drug development repositioning of currently available drugs to new therapeutic indications is considered an alternative route [1]. To date most repositioned drugs have been the consequence of serendipitous observations of unexpected efficacy and side effects of drugs in development or on the market. However, recently, systems biology approaches have been applied in efforts to discover unknown effects for existing drugs.

For instance, drug repositioning approaches have incorporated *in silico* approaches for analyzing large data sets such as gene expression profiles [4,5], literature mining [6], chemical similarity [7], side-effect similarity [8], disease-drug network [9], pathway-based disease network [10], and phenotypic disease network [11]. To establish a more logical approach to repositioning a known drug to a new indication, I established a knowledge platform comprising binary linkages between diseases, drugs, and proteins, from which new and previously unknown connections can be drawn between drugs and diseases of interest. This integrated database was designated PharmDB.

For probing the database and identifying disease-drug linkages, I have developed the Shared Neighborhood Scoring (SNS) algorithm, which predicts relationships between drugs, proteins and diseases. While the relationship data are collected from experiments, coverage of the data is still incomplete. Thus there may be undetected links and hidden nodes in the network. Up to now, a number of prediction methods and measures have been proposed to find these undetected associations from topological or structural properties of various complex networks [12,13]. To date, most of these algorithms and measures are applicable only to a monopartite network that consists only of one type of node. Therefore, multipartite network composed of more

than a type of nodes cannot be analyzed using these measures. To solve this problem, researchers have used projection methods that convert multipartite networks into monopartite ones. Unfortunately, any projection method can result in information loss, especially in low-degree nodes. Accordingly projecting the PharmDB tripartite network into monopartite drug, protein and disease networks can distort many well-known network measures, such as average path length <l>, average clustering coefficient <C>, degree-dependent clustering coefficient C(k), degree distribution P(k), assortativity coefficient r [14], and degree-degree correlation coefficient $k_{nn}(k)$ [15]. To overcome these limits of the projection technique, I designed a new prediction method called Shared Neighborhood Scoring (SNS) algorithm which calculates the probability of a link existence between two nodes of interest. This can be done by evaluating the connections of their neighbors in PharmDB tripartite network.

Results

System overview

The PharmDB is a tripartite pharmacological network database consisting of three kinds of nodes: human diseases, FDA approved drugs or druggable chemicals, and proteins. The proteins in PharmDB include therapeutic targets, disease-associated proteins, and drug-metabolizing proteins. The nodes and links used to construct this network database were imported from nine public databases, namely, EntrezGene interaction [16], MINT [17], DIP [18], CTD [19], TTD [20], ChemBank [21], PharmGKB [22], OMIM [23], and GAD [24] (Table I-1).

Although these individual databases provide information about the relationships between drugs, diseases, and proteins, they do not provide an integrated network map among the three components in an interactive manner. For data integration in a unified format, I adopted PubChem CID for drugs, GeneID for proteins (tagging separate IDs for isozymes and subunits), and MeSH descriptor for diseases (Figure I-1). PharmDB currently includes the nodes of 11,792 drugs, 38,056 proteins, and 6,607 diseases. It also contains 189,800 Drug-Protein, 109,124 Protein-Disease, and 12,232

Drug-Disease, 156,902 Protein-Protein links. The contents of the tripartite pharmacological network in PharmDB are provided through a website (http://www.i-pharm.org/). phExplorer, a graphical data visualization software program is also provided for interactive browsing of relevant data. For constructing workflows, PharmDB is provided in BioMart format (http://biomart.i-pharm.org/). Currently, software for finding the shortest path between two nodes is only provided through the website.

Shared Neighborhood Scoring (SNS) algorithm

The concept of SNS algorithm is similar to Swanson's ABC model, which applies the transitivity rule to discover missing knowledge from biomedical literature [25]. The SNS algorithm is based on the observation that the probability of connection between two nodes shows monotonic increase with "Shared Nodes Count", the number of in-between nodes connecting two nodes (Figure I-2, middle left box). Further the weights for all possible pairs of the network were calculated. First each connected pairs directly linked between two nodes was assigned weight 1. If a pair of two nodes is not connected, the connection probability is assigned as weight for this indirect link or a

virtual link between two nodes. As shown in the Figure I-2, the connection probability for given "Shared Nodes Count" can be computed to be the fraction of directly connected pairs among the total number of pairs having the given "Shared Nodes Count". Finally the share neighborhood score (SN score) was developed by summing up "Shared Nodes Count", the number of shared nodes and "Shared Nodes Weight", the product of each weight of (direct or indirect) links bridging the two end nodes (Figure I-2, bottom left). As the SN score possesses a range of values in each relation category (drug-protein, protein-disease, and drug-disease), a normalization method using the connecting probability function of SN score distribution was developed (see Materials and Methods for details).

The "Shared Nodes Count" distribution for connected pairs and unconnected pairs were compared. Connected pairs shared more neighborhood nodes than unconnected pairs. The p-values of the Kolmogorov-Smirnov (KS) test are less than 2.2e-16 in all three relation categories, meaning that connected pairs and unconnected pairs have significantly distinct distribution (Figure I-3A, I-3B, and I-3C).

The prediction performance of the SNS algorithm was measured by plotting receiver operating characteristic (ROC) curves (Figure I-3D, I-3E, I-3F). For

calculating SN scores, “simple algorithm” considers only “Shared Nodes Count” but “extended algorithm” includes both “Shared Nodes Count” and “Shared Nodes Weight”. As shown in the Figure I-3, the extended algorithm shows better performance than simple one. AUC values of simple algorithm are 0.679, 0.778, and 0.602, in Drug-Protein relation, Drug-Diseae relation, and Protein-Disease relation, respectively. And AUC values for extened algorithm are 0.937, 0.868, and 0.871. According to the result, prediction performances with extended scope of shared neighborhood nodes were improved by 38%, 12%, and 45%, respectively.

Case Study – Benzthiazide as a potential agent for lung cancer

As a case study, squamous cell carcinoma (SCC) (MeSH descriptor: D002294), a subtype of lung cancer, was selected and tested whether PharmDB could identify any drugs that have a potential for treating this type of cancer. For the primary selection of drug candidates in this case, the following criteria was made. First, they should be inferred by SNS algorithm with SN score bigger than 0.004 and Share Nodes Count zero. Second, they should belong to FDA approved drugs. Third, they should not have been previously used for cancer drug. Forth, they should be directly linked to cancer

target proteins (Figure I-4). Twenty eight common drugs fit to the four criteria above and were suggested as potential SCC drug candidates (Supplementary Table I-3). I then went over these candidates to choose the one for experimental validation. Considering technical feasibility, availability of materials, intellectual property and potential for new drug development, I decided to examine thia-benzthiazide (TBZT) whether it can be used for SCC treatment. TBZT is a kind of thiazide diurectic used for the treatment of high blood pressure and edema [26]. To validate a potential of TBZT as a lung cancer drug, different concentrations of TBZT were administered to squamous lung cancer cells (HCC-1588) under hypoxic conditions (which mimic the tumor microenvironment), as well as under normoxic conditions [27]. Their effects on cell proliferation were monitored by [^{3}H] thymidine incorporation. Under hypoxic conditions only, TBZT can suppress proliferation of cancer cell in a dose-dependent manner (Figure I-5A). The hypoxia-dependent cell death induced by TBZT was further confirmed by flow cytometry (Figure I-5B).

Carbonic anhydrases (CAs) are zinc metalloenzymes which catalyze the conversion of carbon dioxide to the bicarbonate ion and protons. The CAs are involved in many biological and physical processes including pH homeostasis and have 16

mammalian isoforms (CA1 ~ CA 16) [28]. In PharmDB, TBZT is linked to carbonic anhydrase 2 (CA2). However, TBZT can suppress proliferation of lung cancer cell under hypoxic conditions only (Figure I-5) and the expression of CA2 is not associated with hypoxic conditions. So I have extended cancer-linked CA isoforms in PharmDB (Figure I-6A). As a result, I considered three different human CA isozymes (i.e., 1, 2, and 9) as targets of TBZT, and tested whether TBZT inhibits CA activity. TBZT suppressed all of the three CA isozymes with similar K_i values (Figure I-6B). As a positive control, acetazolamide (AZA), a known inhibitor of carbonic anhydrases (CAs), was also used [29]. AZA also suppressed the activities of the three CAs, although the K_i values varied depending on the target enzymes. However, among the CA isozymes, CA9 is known to be induced in hypoxic conditions and has functional association with cancer [30]. Thus, the efficacy of TBZT against HCC-1588 cells is likely to have resulted from its inhibition of CA9. For that reasons, I decided to focus on CA9 as the major effective target of TBZT against cancer although I do not exclude the involvement of other isozymes.

CA9, a carbonic anhydrase isoenzyme, is a transmembrane protein that plays an important role in pH regulation [31]. The expression of CA9 is highly induced in

various cancers under hypoxic conditions, which is functionally important for the growth and survival of tumor cells [31]. I confirmed whether CA9 is actually induced in hypoxic conditions by Western blotting with its specific antibody in HCC-1588. As expected, CA9 levels were significantly increased in hypoxic conditions (1% O_2) compared with those in normoxic conditions (20% O_2) (Figure I-6C). I also confirmed that TBZT induced cell death by measuring the activation of caspase 3 (Figure I-6D). To confirm the drug-protein pair relationship between CA9 and TBZT, I tested whether the forced expression of CA9 would compensate for the anti-proliferative activity of CA9 by the treatment of TBZT under hypoxic conditions. Cell proliferation was reduced to 70% of the control cells by the treatment of TBZT in the cells transfected with EV, but 35% of the control cells in the cells transfected with CA9. Therefore, the exogenous supplementation of CA9 recovered the proliferation by up to 35% (Figure I-6E). This result validates that the anti-proliferative activity of TBZT against HCC-1588 cells mainly involves CA9. Perhaps, the remaining part could be contributed by other CA isozymes that are also involved in the regulation of cancer. Even if further chemical optimization of TBZT is required to improve efficacy and specificity, these results suggest a possible application of TBZT for further development against lung cancer

through its CA9 inhibitory activity.

Figure I-1. Overview of PharmDB

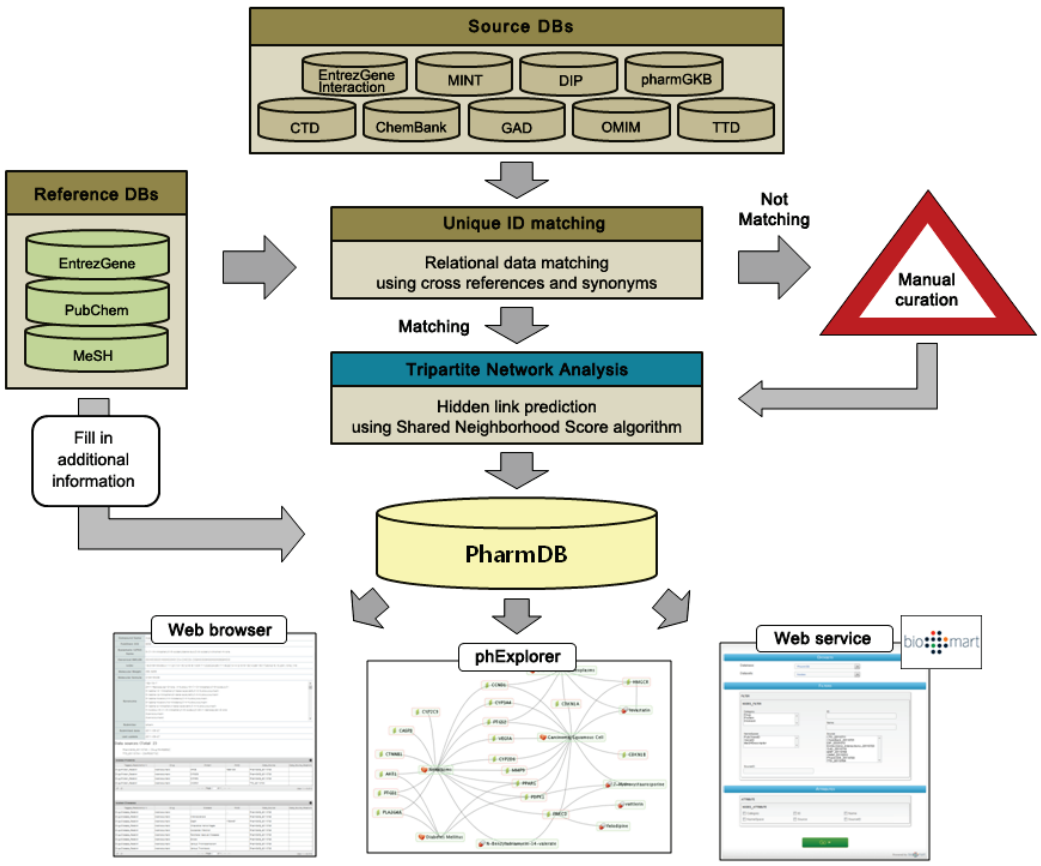


Figure I-1. Overview of PharmDB. Nine different databases were integrated using standard IDs (Entrez Gene ID for protein, PubChem CID for drug and MeSH Descriptor ID for disease) to construct PharmDB. The integrated network was analyzed using the shared neighborhood scoring algorithm, providing a predictive capacity for PharmDB to suggest functional relationships between diseases, proteins, and rugs. These data are provided through a web browser, phExplorer (network visualization software) and web service (http://www.i-pharm.org/, http://biomart.i-pharm.org/).

Figure I-2. Shared neighborhood scoring algorithm

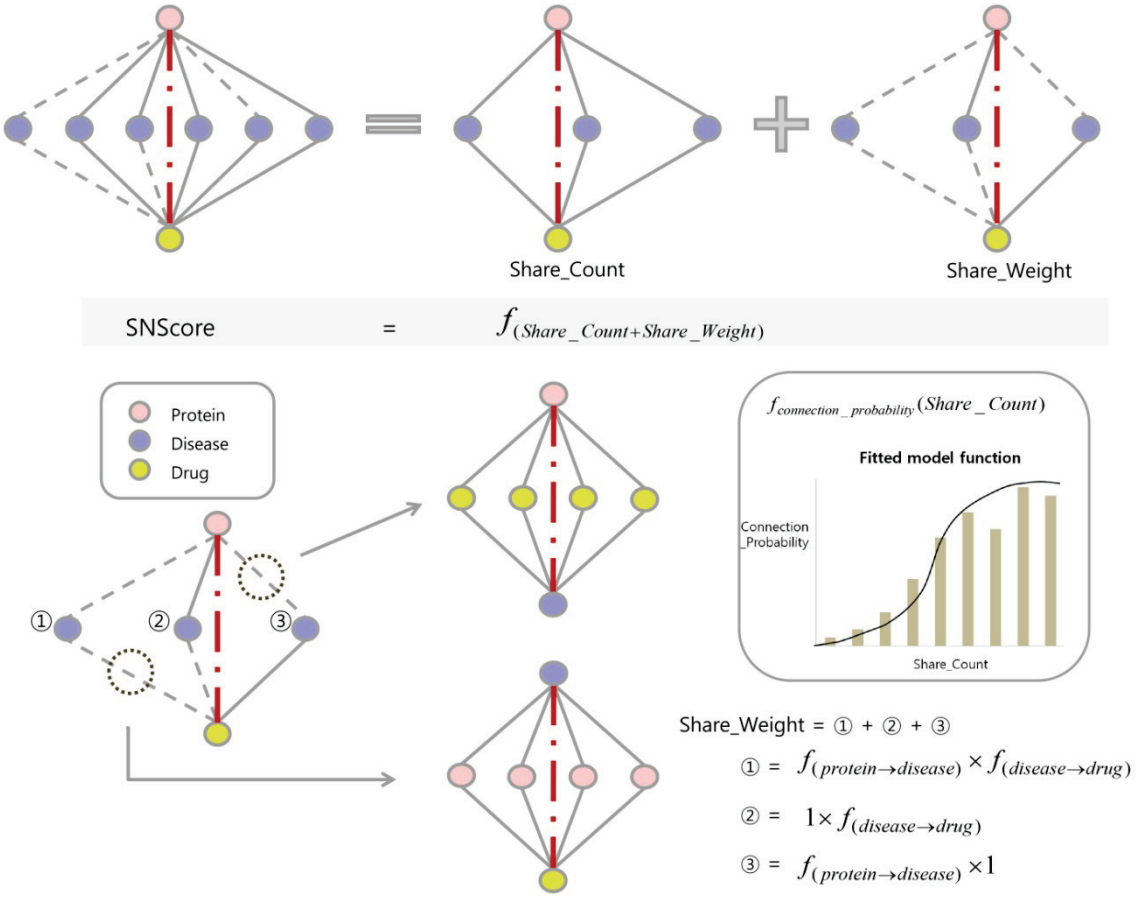


Figure I-2. Shared neighborhood scoring algorithm. SN score can be calculated by summing up "Shared Nodes Count" and "Shared Nodes Weight". First, “Shared Nodes Count” is defined as the number of shared nodes to consider the effect of direct connectors. Similary, “Shared Nodes Weight” is defined as the product of each weight of links bridging two end nodes to trace the effect of indirect neighbors (Bottom right). Here weight is a measure for connecting probability of each pair. For all possible pairs of the network, firstly weight 1 is assigned to each connected pairs directly linked between two nodes. Weight for unconnected pairs is assigned the connection probability, the fraction of directly connected pairs among the total number of pairs having the given “Shared Nodes Count”. For example, (1) the weight product of two indirect links (2), (3) weight of the upper (direct link / indirect link) multiplied by weight of the lower one (indirect link / direct link). The sum of (1), (2), and (3) is "Shared Weight" of the unconnected pair (protein, drug) (bottom left).

Figure I-3. Shared neighborhood node distribution and evaluation of the shared neighborhood scoring algorithm

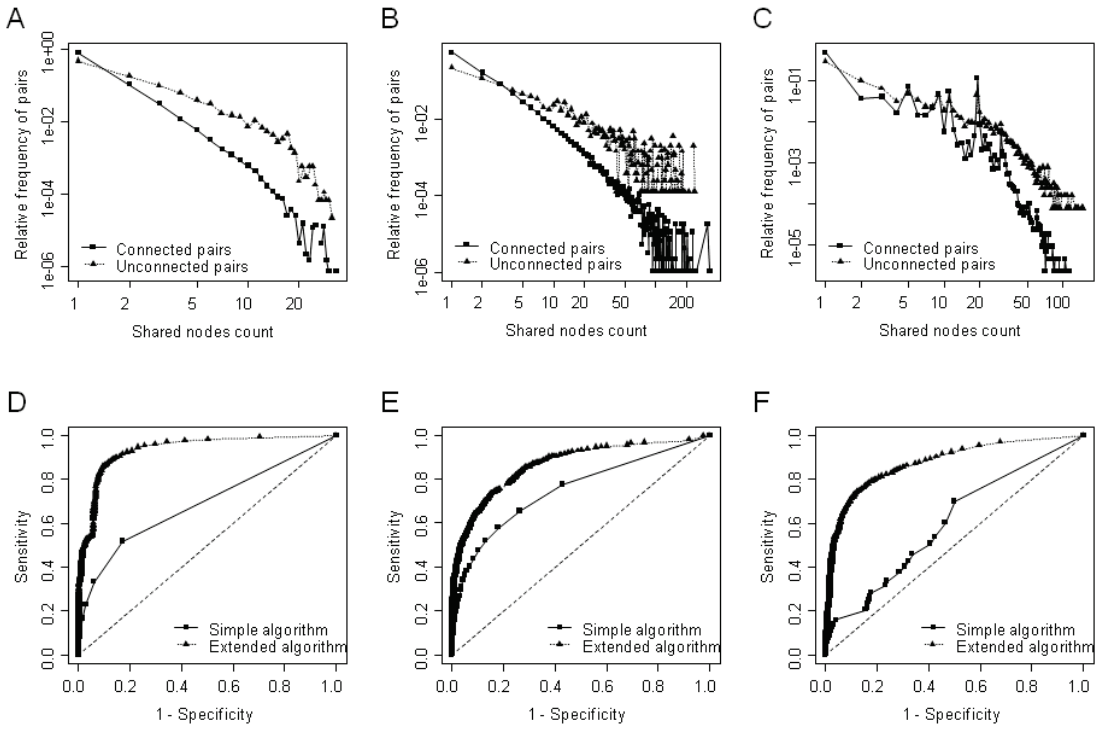


Figure I-3. Shared neighborhood node distribution and evaluation of the shared neighborhood scoring algorithm. Shared neighborhood node distribution comparison between connected links and unconnected links in Drug-Protein relation (A), Drug-Disease relation (B) and Protein-Disease relation (C) (Rectangle: Connected links, Triangle: Unconnected links). ROC analysis of simple form of SNS algorithm and extended form of SNS algorithm in Drug-Protein relation (D), Drug-Disease relation (E) and Protein-Disease relation (F) (Rectangle: Simple algorithm, Triangle: Extended algorithm).

Figure I-4. Drug repositioning pipeline overview

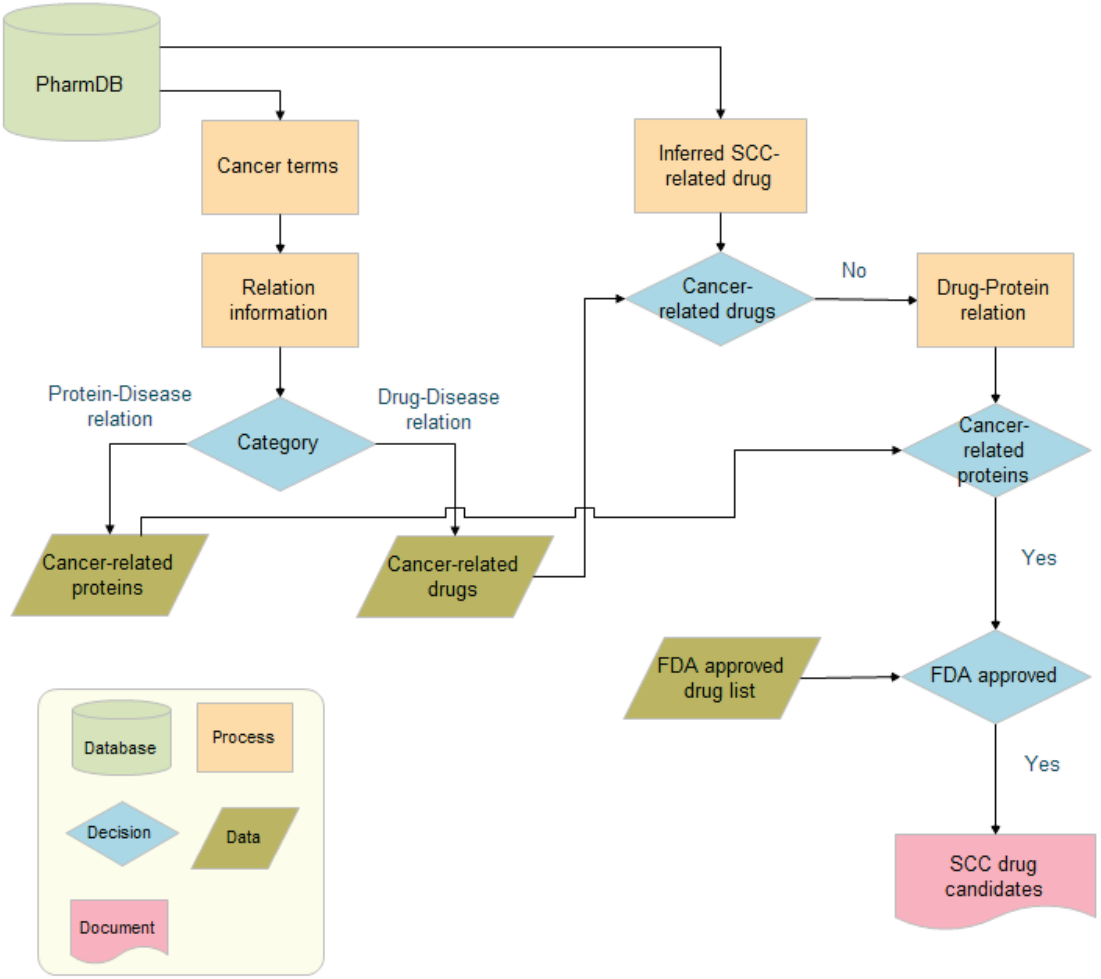


Figure I-4. Drug repositioning pipeline overview. Schematic representation of drug repositioning pipeline for squamous cell carcinoma (SCC). First, cancer-related proteins and drugs were extracted from PharmDB using cancer terms (such as “Carcinoma”, “Neoplasm”, and “Cancer”). Second, inferred SCC-related drugs were extracted using the shared neighborhood scoring algorithm. Among the candidates, any known cancer agents were filtered out; leaving only drugs that had not been previously implicated as anti-cancer drugs. Then the FDA approved drugs which known to be related with cancer-related proteins were maintained for further analysis as SCC drug candidates in this study.

Figure I-5. The hypoxia-dependent TBZT effect against SCC

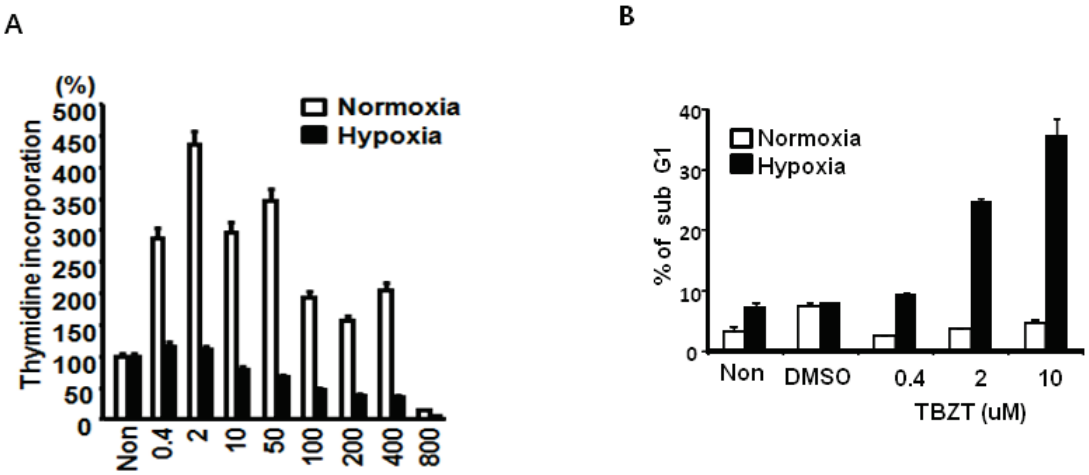


Figure I-5. The hypoxia-dependent TBZT effect against SCC. (A) Antiproliferative activity of TBZT was monitored by [^{3}H] thymidine incorporation under normoxic and hypoxic conditions. (B) The effect of TBZT on cell death was monitored by counting sub-G1 cells

Figure I-6. TBZT as an inhibitor of CA9

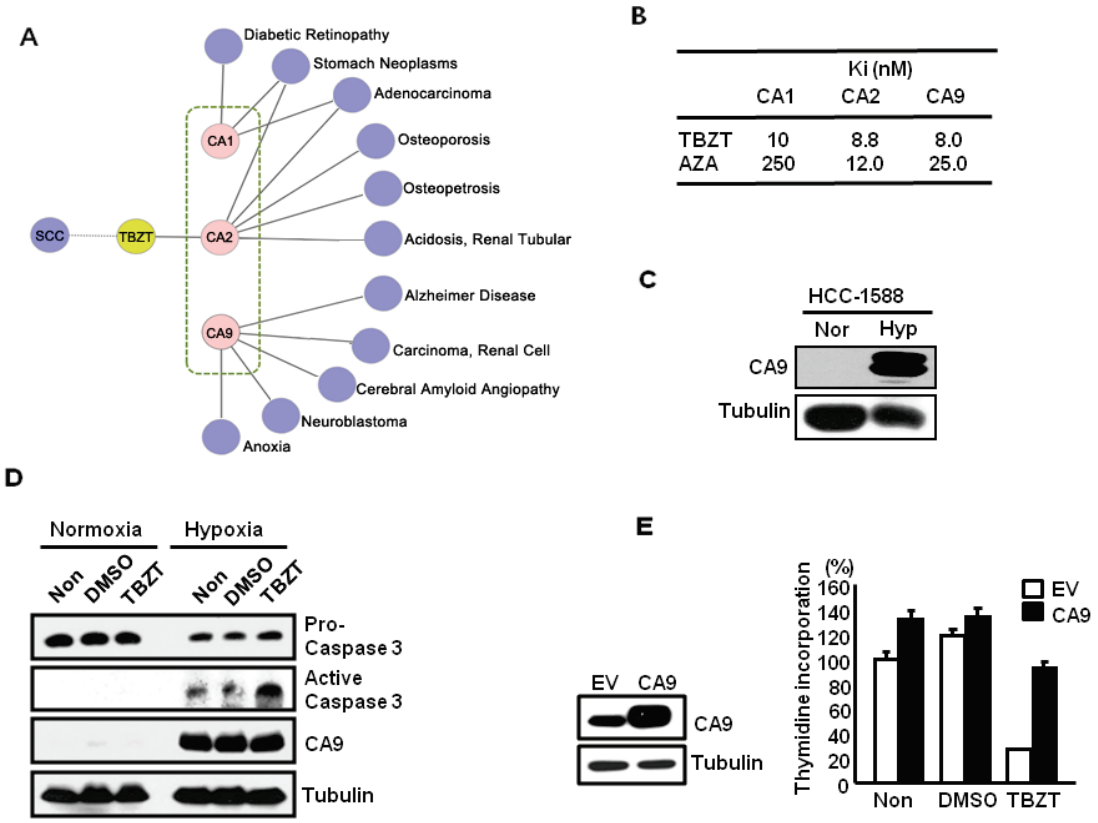


Figure I-6. TBZT as an inhibitor of CA9. (A) To validate predictions by PharmDB analysis for SCC, TBZT is tested for inhibitory activity against its potential targets, CA isozymes (CA1, CA2, and CA9). (B) *In vitro* inhibition of TBZT and the AZA control against CA isoforms (i.e., 1, 2, and 9). (C) Cellular levels of CA9 in the SCC cell line, HCC-1588, under normoxic and hypoxic conditions. (D) The effect of TBZT on cell death was monitored by caspase-3 activation. (E) HCC-1588 cells, transfected with an empty vector (EV) or CA9, were treated with TBZT under normoxic and hypoxic conditions.

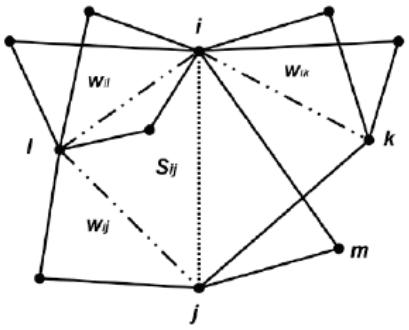
Table I-1. Data sources of PharmDB

	Drug-Protein Relation	Protein-Disease Relation	Drug-Disease Relation	Protein-Protein Interaction	Update Date
EntrezGene Interaction				V	2011.07.20
MINT				V	2011.07.08
DIP				V	2010.10.10
PharmGKB	V	V	V	V	2011.07.20
CTD	V	V	V		2011.07.11
TTD	V	V	V		2011.07.04
ChemBank			V		2011.07.21
OMIM		V			2011.03.10
GAD		V			2011.07.16

Supplementary Figure I-1. Design of shared neighborhood score algorithm

SN Score = $f_{(Share\_Count+Share\_Weight)}$

Share weight = Sum of connecting probability



Share weight

$$S_{ij} = S_{ij}^{(0)} + S_{ij}^{(1)} + S_{ij}^{(2)}$$
$$= \sum_k (a_{ik} + P(S_{ik}^{(0)})\delta_{a_{ik},0})(a_{kj} + P(S_{kj}^{(0)})\delta_{a_{kj},0})$$

where,

$S_{ij}^{(0)} = \sum_k a_{ik}a_{kj}$

$S_{ij}^{(1)} = \sum_k (a_{ik}P(S_{kj}^{(0)})\delta_{a_{kj},0} + a_{kj}P(S_{ik}^{(0)})\delta_{a_{ik},0})$

$S_{ij}^{(2)} = \sum_k P(S_{ik}^{(0)})\delta_{a_{ik},0}P(S_{kj}^{(0)})\delta_{a_{kj},0})$

a_{ik}=1 and $\delta_{a_{ik},0}$=0 if the link between i and k is real

a_{ik}=0 and $\delta_{a_{ik},0}$=1 if the link is virtual

P(x) = connecting probability function

$$P_{(x)} = \frac{(\text{number of connected pairs})}{(\text{number of pairs})} \text{ for n} = 1,2,\ldots$$

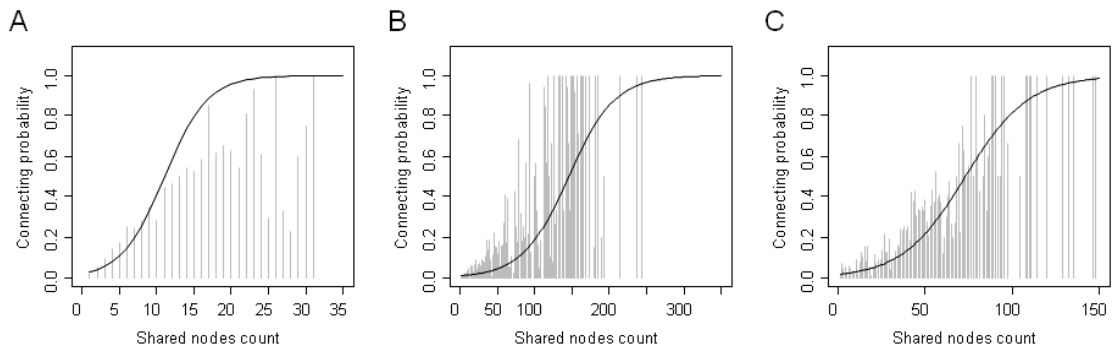
$$P_{(x)} = 0 \text{ for } x = 0$$

Logistic function is used for non-linear regression.

$$f(x) = \frac{1}{1+e^{a+bx}}$$

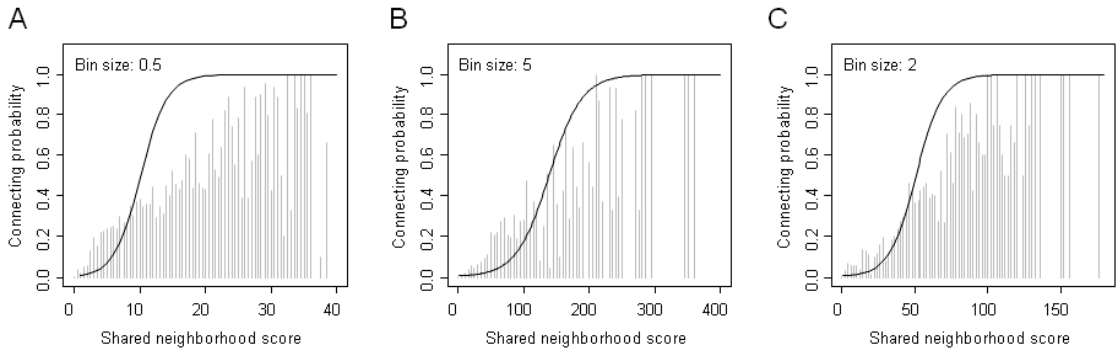
Supplementary Figure I-1. Design of shared neighborhood score algorithm. The shared neighborhood scoring algorithm is based on the basic principle that the connection probability of a link between two nodes (*i* and *j*) is roughly proportional to the number of nodes commonly shared between the original two nodes, *i* and *j*.

Supplementary Figure I-2. Non-linear regression results for extracting connecting probability function



Supplementary Figure I-2. Non-linear regression results for extracting connecting probability function. Non-linear connecting probability model based on the number of nodes commonly shared between two nodes. (A) Drug-Protein. (B) Drug-Disease. (C) Protein-Disease.

Supplementary Figure I-3. Connecting probability model for SN score



Supplementary Figure I-3. Connecting probability model for SN score. The shared neighborhood score is proportional to the number of shared neighborhood nodes. As the amount of data is not evenly distributed on each relation category, even if two different types of relations have identical score, their connecting possibility can't be regarded as identical. Therefore the shared neighborhood score is normalized using connecting probability function. (A) Drug-Protein. (B) Drug-Disease. (C) Protein-Disease.

Supplementary Table I-1. Model coefficient values from non-linear regression for connecting probability function

Model Coefficient	Drug-Protein	Drug-Disease	Protein-Disease
a	3.87	4.67	4.03
b	-0.35	-0.03	-0.06

Supplementary Table I-2. Model coefficient values from connecting probability model for SN score

Model Coefficient	Drug-Protein	Drug-Disease	Protein-Disease
a	5.09	5.62	5.53
b	-0.50	-0.11	-0.04

Supplementary Table I-3. Inferred SCC drug candidates

Twenty eight SCC drug candidates which satisfy four criteria; 1) inferred by SNS algorithm 2) FDA approved drugs 3) non-cancer drugs 4) directly linked to cancer-related proteins.

Name	Cancer-related GeneID	Cancer-related GeneSymbol
MIRTAZAPINE	3157	HMGCS1
ZONISAMIDE	7364	UGT2B7
ARGATROBAN	2147	F2
GANCICLOVIR	6580	SLC22A1
ACYCLOVIR	6580	SLC22A1
CLADRIBINE	9154	SLC28A1
BENZTHIAZIDE	760	CA2
HYDROCHLOROTHIAZIDE	760	CA2
CHLOROTHIAZIDE	760	CA2
DIAZOXIDE	3767	KCNJ11
DROPERIDOL	3757	KCNH2
FLURBIPROFEN	8856	NR1I2
GLIMEPIRIDE	8856	NR1I2
GLIPIZIDE	1559	CYP2C9
METHYLDOPA	1312	COMT
MEPROBAMATE	5581	PRKCE
MYCOPHENOLATE MOFETIL	54576	UGT1A8
AZITHROMYCIN	1813	DRD2
PINDOLOL	155	ADRB3
FLUPHENAZINE DECANOATE	3351	HTR1B
PROCHLORPERAZINE	1813	DRD2
LATANOPROST	5737	PTGFR
NALOXONE	8856	NR1I2
NALBUPHINE	4985	OPRD1
CABERGOLINE	1813	DRD2
TOBRAMYCIN	4549	RNR1
ZAFIRLUKAST	10800	CYSLTR1
NATEGLINIDE	1559	CYP2C9

Discussion

This study demonstrates that drug repositioning can be rapidly guided by a knowledge platform PharmDB, a pharmacological network database comprising protein, drug, and disease data which are freely available as a web-based service. As an ever-increasing amount of biological and pharmacological data are scattered throughout the literature and in proprietary databases, the integrated data of PharmDB provides a valuable tool by consolidating certain valuable sets of data. I adopted a tripartite pharmacological network-based analysis, and developed a novel neighborhood scoring algorithm to predict previously unknown relationships between drugs, proteins and diseases. The theoretical foundation of algorithm is that a connection probability between two nodes is proportional to the number of nodes commonly shared between them. So the connection probability of two indirectly linked nodes was computed, which is called the shared neighborhood score. This score can highlight missing linkages which may either result from “no actual connection” or “lack of information” and help to differentiate between these two possibilities.

I experimentally validated the usefulness of the shared neighborhood score by

identifying a hitherto unknown drug-protein relationship and potential new indication based on this connection. Aside from drug repositioning, the network map of PharmDB composed of not only the data integrated from diverse databases but also the predicted data using the shared neighborhood algorithm can applied to other purposes, such as the prediction of drug mode-of-action, off-target effects, and even the design of optimal drug combinations for a disease of interest.

PharmDB, an integrated tripartite database, coupled with Shared Neighborhood Scoring (SNS) algorithm, would provide much more enriched information than general integrated databases and give us clues for finding new indication of known drugs. Furthermore, these data can be easily explored with phExplorer and accessed via BioMart web service (http://www.i-pharm.org/, http://biomart.i-pharm.org/).

Materials and methods

Construction of PharmDB

To integrate the data in the existing databases that contain different identifiers, I assigned the following standard identifiers (IDs): PubChem CID for drug, GeneID for protein, and MeSH descriptor for disease. A comprehensive drug-protein-disease tripartite network was constructed by integrating the link information from the nine databases, namely, EntrezGene interaction (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/), MINT (ftp://mint.bio.uniroma2.it/pub/release/mitab26/2011-07-08/), CTD (http://ctd.mdibl.org/), TTD (http://bidd.nus.edu.sg/group/cjttd/TTD_Download.asp), ChemBank (http://chembank.broadinstitute.org), DIP (http://dip.doe-mbi.ucla.edu/dip/Download.cgi), PharmGKB (http://www.pharmgkb.org/resources/downloads_and_web_services.jsp), OMIM (ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/), and GAD (http://geneticassociationdb.nih.gov/). As the existing databases have their own unique ID systems, I tagged the standard IDs using an in-house script. For the entities that were not tagged by the in-house script, I manually assigned them with appropriate IDs.

Shared neighborhood scoring algorithm

The shared neighborhood scoring algorithm is based on the basic principle that the connection probability of a link between two nodes (*i* and *j*) is roughly proportional to the number of nodes commonly shared between the original two nodes, *i* and *j* (Supplementary Figure I-1). The shared neighborhood score S_{ij} is defined as $S_{ij} = \sum_k W_{ik} W_{kj}$. In this equation, *i* and *j* indicate the indices of a pair of nodes; *k* is the index of a shared neighbor node; and W_{ik} is the weight of a link between *i* and *k*. The link between *i* (or *j*) and *k* can be real or virtual (i.e., having no known connection but is expected to be connected). Thus, I can define W_{ik} as $W_{ik}=a_{ik}+P\left(S_{ik}^{(0)}\right)\delta_{a_{ik},0}$. Here, $a_{ik}=1$ and $\delta_{a_{ik}0}=0$ if the link between *i* and *k* is real; and $a_{ik}=0$ and $\delta_{a_{ik}}=1$ if the link is virtual. When there are only direct connections between node *i* and node *j*, a 0th-order shared neighborhood score $S_{ij}^{(0)}$ becomes $S_{ij}^{(0)} = \sum_k a_{ik} a_{kj} = n(0,1,2,3...)$, where n is the number of bridging nodes between node *i* and node *j*. $P(S_{ik}^{(0)})$ is a connection probability that depends on the value of the 0th-order shared neighborhood score $S_{ik}^{(0)}$. For the 0th-order shared neighborhood score $S_{ik}^{(0)} = n(=0,1,2,...)$, the function $P(n)$ is defined as follows:

$$P(n)=\frac{(number\ of\ connected\ pairs)}{(number\ of\ pairs)}\ for\ n=1,2,p$$

$P(n)=0$ *for* $n=0$

Based on the probability above, non-linear regression was carried out to extract connecting probability functions. Logistic function was used for this (Supplementary Figure I-2, Supplementary Table I-1).

$$f(x)=\frac{1}{1+e^{a+bx}}$$

The shared neighborhood score s_{ij} then becomes

$$s_{ij}=s_{ij}^{(0)}+s_{ij}^{(1)}+s_{ij}^{(2)}=\sum_k (a_{ik}+P(S_{ik}^{(0)})\delta_{a_{ik}0})(a_{kj}+P(S_{kj}^{(0)})\delta_{a_{kj}0})$$

where $S_{ij}^{(0)}=\sum_k a_{ik}a_{kj}$, $S_{ij}^{(1)}=\sum_k (a_{ik}P(S_{kj}^{(0)})\delta a_{kj}0+a_{kj}P(S_{ik}^{(0)})\delta_{a_{ik}0})$, and $S_{ij}^{(2)}=P(S_{ik}^{(0)})\delta_{a_{ik}0}P(S_{kj}^{(0)})\delta_{a_{kj}0}$. Here, the 1st-order term $S_{ij}^{(1)}$ is added when some nodes are linked directly to node *i* (or *j*) but linked indirectly to node *j* (or *i*). The 2nd-order term $S_{ij}^{(2)}$ is considered only when some nodes are linked indirectly to both node *i* and node *j*. On Supplementary Figure I-1, as node *m* is the only shared neighbor of node *i* and node *j*, $S_{ij}^{(0)}=1$. To obtain the 1st-order shared neighborhood score $S_{ij}^{(1)}$ or the 2nd-order shared neighborhood score $S_{ij}^{(2)}$, connection probability $P(S_{ik}^{(0)})$ is calculated beforehand. As a pair (*i*, *k*) is mediated by two nodes, $W_{ik}=P(2)$, Path (*i*, *k*, *j*) is composed of an indirect link (*i*, *k*) and a direct link (*k*, *j*). Similarly, $W_{il}=P(3)$ and $W_{lj}=P(1)$. Path (*i*, *l*, *j*) is composed of both indirect links (*i*, *l*) and (*l*, *j*). The total

shared neighborhood score is thus $s_{ij} = S_{ij}^{(0)} + S_{ij}^{(1)} + S_{ij}^{(2)} = 1 + P(2) + P(3)P(1)$. When calculating $S_{ij}^{(2)}$, I omitted a link between node *i* and *j* to remove the dependency of the measure on the existence of a link between *i* and *j,* which is the so-called "leave-one-out approach"[32].

The shared neighborhood score is proportional to the number of shared neighborhood nodes. So there is no an upper limit on score. The problem is that the amount of data is not evenly distributed on each relation category. So, even if two different types of relations have identical score, their connecting possibility can't be regarded as identical. For that reason, I normalized the shared neighborhood score using connecting probability function (Supplementary Figure I-3).

FDA approved drugs

I have downloaded Drugs@FDA data files (Last updated: 19/09/2011)(http://www.fda.gov/downloads/Drugs/InformationOnDrugs/UCM163762). Then I extracted single active ingredient from Product table and tagged PubChem ID for them. The total number of FDA approved drugs tagged with PubChem ID is 23,191.

Cell culture and materials

The HCC-1588 cell line was obtained from the Korean cell line bank and was maintained in RPMI (Hyclone) containing 10% fetal bovine serum and 1% antibiotics. Antibody against caspase-3 and tublin (Cell Signaling Technology) were purchased. M73 monoclonal antibody to CA9 was obtained from Dr. S. Pastorekova (Slovak Academy of Science, Slovak Republic). TBZT and AZA were purchased from Sigma.

Thymidine incorporation assay

To determine the effect of TBZT on cell proliferation, HCC-1588 cells were treated with TBZT in 2% serum-containing media for 48 h under normoxic (20% O_2) and hypoxic (1% O_2) conditions. AZA was used as positive control. pcDNA3-CA9 vector and empty vector (Dr. J.-Y. Kim, National Cancer Center, Korea) were transfected into HCC-1588 cells using Lipofectamine 2000 (Invitrogen). After 24 h incubation, TBZT was added to 2% serum-containing media for 48 h under hypoxic conditions. [^{3}H] thymidine at 1 μCi/ml was added to the culture medium and was incubated for 4 h. The incorporated thymidine was measured by liquid scintillation counter (Wallac).

Flow cytometry

HCC-1588 cells were treated with TBZT (0.4, 2, 10 μM) in 2% serum -containing medium for 48 h under normoxic and hypoxic conditions. AZA was used as positive control. The treated cells were fixed with 70% ethanol for 1 h at 4°C, washed twice with ice-cold PBS, and stained with propidium iodide (50 μg/ml) containing 0.1% sodium citrate, 0.3% NP-40 (nonylphenoxylpolyethoxylethanol 40), and 50 μg/ml RNase A for 40 min. The cells were subjected to flow cytometry (FACSCalibur, Becton-Dickinson) to evaluate the apoptotic cells by counting the sub-G1 cells. For each sample, 20,000 cells were analyzed using Cell Quest Pro software.

Enzyme activity

An applied photophysics stopped-flow instrument was used for assaying CA-catalyzed CO_2 hydration activity [32]. Following the initial rates of the CA-catalyzed CO_2 hydration reaction for a period of 10–100 s, phenol red (at a concentration of 0.2 mM) was used as the indicator, working at the absorbance maximum of 557 nm in 20 mM HEPES buffer (pH 7.5) and 20 mM Na_2SO_4 (to maintain the constant ionic strength). For the determination of the kinetic parameters and inhibition constants, the CO_2

concentrations used ranged from 1.7–17 mM. For each inhibitor, at least six traces of the initial 5–10% of the reaction were used for determining the initial velocity. The uncatalyzed rates were determined in the same manner and were subtracted from the total observed rates. Stock solutions of the inhibitor (0.1 mM) were prepared in distilled-deionized water and diluted to 0.01 nM with distilled-deionized water. Inhibitor and enzyme solutions were preincubated together for 15 min–72 h at room temperature (15 min) or 4°C (all other incubation times) prior to assay to allow the formation of the enzyme-inhibitor complex or the eventual active site mediated hydrolysis of the inhibitor. The inhibition constants were obtained by non-linear least-squares methods using PRISM 3 as previously described. The mean values were represented from at least three different determinations [31,33].

Acknowledgments

This study was supported by the grants of the Global Frontier (NRF-M1AXA002-2010-0029785) and the Research Information Center Supporting Program (2012-0000350) and the WCU project (R31-2008-000-10103-0) of the Ministry of Education, Science, and Technology and Korea Healthcare Technology (A092255-0911-1110100), the Ministry of Health and Welfare Affairs, and Gyonggi-do to SK, an EU project of the 7th framework programme (METOXIA) to CTS, and by the Korean Ministry of Education, Science and Technology (MEST) under grant number 20110002321.

References

1. Loging W, Harland L, Williams-Jones B: **High-throughput electronic biology: mining information for drug discovery**. *Nature reviews Drug discovery* 2007, **6**(3):220-230.

2. Butcher EC, Berg EL, Kunkel EJ: **Systems biology in drug discovery**. *Nature biotechnology* 2004, **22**(10):1253-1259.

3. Wishart DS: **Discovering drug targets through the web**. *Comparative biochemistry and physiology Part D, Genomics & proteomics* 2007, **2**(1):9-17.

4. Lamb J: **The Connectivity Map: a new tool for biomedical research**. *Nature reviews Cancer* 2007, **7**(1):54-60.

5. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN *et al*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**(5795):1929-1935.

6. Li J, Zhu X, Chen JY: **Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts**. *PLoS*

computational biology 2009, **5**(7):e1000450.

7. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB *et al*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**(7270):175-181.

8. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity**. *Science* 2008, **321**(5886):263-266.

9. Hu G, Agarwal P: **Human disease-drug network based on genomic expression profiles**. *PloS one* 2009, **4**(8):e6536.

10. Li Y, Agarwal P: **A pathway-based view of human diseases and disease relationships**. *PloS one* 2009, **4**(2):e4346.

11. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA: **A dynamic network approach for the study of human phenotypes**. *PLoS computational biology* 2009, **5**(4):e1000353.

12. Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, Francois JM, Zecchina R: **Finding undetected protein associations in cell signaling by belief propagation**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(2):882-887.

13. Lu L, Jin CH, Zhou T: **Similarity index based on local paths for link prediction of complex networks**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2009, **80**(4 Pt 2):046122.

14. Newman ME: **Mixing patterns in networks**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2003, **67**(2 Pt 2):026126.

15. Pastor-Satorras R, Vazquez A, Vespignani A: **Dynamical and correlation properties of the internet**. *Physical review letters* 2001, **87**(25):258701.

16. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic acids research* 2011, **39**(Database issue):D52-57.

17. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update**. *Nucleic acids research* 2010, **38**(Database issue):D532-539.

18. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic acids research* 2004, **32**(Database issue):D449-451.

19. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ: **The Comparative Toxicogenomics Database: update**

2011. *Nucleic acids research* 2011, **39**(Database issue):D1067-1072.

20. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C *et al*: **Update of TTD: Therapeutic Target Database**. *Nucleic acids research* 2010, **38**(Database issue):D787-791.

21. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M *et al*: **ChemBank: a small-molecule screening and cheminformatics resource database**. *Nucleic acids research* 2008, **36**(Database issue):D351-359.

22. Thorn CF, Klein TE, Altman RB: **Pharmacogenomics and bioinformatics: PharmGKB**. *Pharmacogenomics* 2010, **11**(4):501-505.

23. **OMIM (TM)**. In. McKusick-Nathans Institute of Genetic Medicine Online Mendelian Inheritance in Man, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD); 2009: http://www.ncbi.nlm.nih.gov/omim/.

24. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nature genetics* 2004, **36**(5):431-432.

25. Swanson DR, Smalheiser NR: **An interactive system for finding complementary**

literatures: a stimulus to scientific discovery. *Artificial intelligence* 1997, **91**(2):183-203.

26. Havard CW, Wood PH: **Clinical evaluation of benzthiazide, an oral diuretic**. *British medical journal* 1960, **1**(5188):1773-1776.

27. Yotnda P, Wu D, Swanson AM: **Hypoxic tumors and their effect on immune cells and cancer therapy**. *Methods Mol Biol* 2010, **651**:1-29.

28. Robertson N, Potter C, Harris AL: **Role of carbonic anhydrase IX in human tumor cell growth, survival, and invasion**. *Cancer research* 2004, **64**(17):6160-6165.

29. Xiang Y, Ma B, Li T, Yu HM, Li XJ: **Acetazolamide suppresses tumor metastasis and related protein expression in mice bearing Lewis lung carcinoma**. *Acta pharmacologica Sinica* 2002, **23**(8):745-751.

30. Winum JY, Rami M, Scozzafava A, Montero JL, Supuran C: **Carbonic anhydrase IX: a new druggable target for the design of antitumor agents**. *Medicinal research reviews* 2008, **28**(3):445-463.

31. Supuran CT: **Carbonic anhydrases: novel therapeutic applications for inhibitors and activators**. *Nature reviews Drug discovery* 2008, **7**(2):168-181.

32. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(8):4372-4376.

33. Khalifah RG: **The carbon dioxide hydration activity of carbonic anhydrase. I. Stop-flow kinetic studies on the native human isoenzymes B and C**. *The Journal of biological chemistry* 1971, **246**(8):2561-2573.

34. Maresca A, Temperini C, Vu H, Pham NB, Poulsen SA, Scozzafava A, Quinn RJ, Supuran CT: **Non-zinc mediated inhibition of carbonic anhydrases: coumarins are a new class of suicide inhibitors**. *Journal of the American Chemical Society* 2009, **131**(8):3057-3062.

Chapter Ⅱ

CDA: Combinatorial drug discovery using transcriptional response modules

Running title: Synergistic combinatorial drug discovery

Keywords: Combinatorial drug, signaling pathway, gene expression

Abbreviations list

CDA: Combinatorial drug assembler

LPA: Lysophosphatidic acid

CDK: Cyclin-dependent kinase

HDACis: Histone deacetylase inhibitors

TNBC: Triple-negative breast cancer

CI: Combination index

DRI: Dose reduction index

KS: Kolmogorov-Smirnov

ES: Enrichment score

Abstract

Anticancer therapies that target single signal transduction pathways often fail to prevent proliferation of cancer cells because of overlapping functions and cross-talk between different signaling pathways. Recent research has identified that balanced multi-component therapies might be more efficacious than highly specific single component therapies in certain cases. Ideally, synergistic combinations can provide 1) increased efficacy of the therapeutic effect 2) reduced toxicity as a result of decreased dosage providing equivalent or increased efficacy 3) the avoidance or delayed onset of drug resistance. Therefore, the interest in combinatorial drug discovery based on systems-oriented approaches has been increasing steadily in recent years.

Here I describe the development of Combinatorial Drug Assembler (CDA), a genomics and bioinformatics system, whereby using gene expression profiling, multiple signaling pathways are targeted for combinatorial drug discovery. CDA performs expression pattern matching of signaling pathway components to compare genes expressed in an input cell line (or patient sample data), with expression patterns in cell lines treated with different small molecules. Then it detects best pattern matching

combinatorial drug pairs across the input gene set-related signaling pathways to detect where gene expression patterns overlap and those predicted drug pairs could likely be applied as combination therapy. I carried out *in vitro* validations on non-small cell lung cancer cells and triple-negative breast cancer (TNBC) cells. I found two combinatorial drug pairs that showed synergistic effect on lung cancer cells. Furthermore, I also observed that halofantrine and vinblastine were synergistic on TNBC cells.

Introduction

Advances in *in vitro* test systems have shifted drug research from animal studies to target-oriented research [1]. Combining this process with genomic research, agents specifically targeting unique proteins related to specific disease have been found. Amongst these successful stories of targeted agents is the BCR-ABL kinase inhibitor imatinib (Gleevec; Novartis), which is using for the treatment of chronic myelogenous leukemia (CML). However, in such cases, drug resistance arises possibly owing to the diversity of mutations of the gene encoding BCR-ABL as well as other pathways on parallel signalling pathways [2]. Despite successes such as these, many other drug candidates targeting disease-associated gene products have been found to be inefficient or to cause severe side effects. So the limitations of the single protein targeted agent paradigm have come to surface.

Living systems rely on complex signaling pathways to maintain their performance in the face of various perturbations [3]. This complexity appears to pose a barrier for anticancer therapies targeting single signalling pathways. Cancer cells possess compensatory mechanisms to overcome perturbations where they occur at one

signalling axis and so therapies targeting only one pathway can fail in clinical trials due to lack of efficacy, or be overcome by mutations at an important receptor [4]. Recent research has identified that in some cases, balanced multi-component therapies might be better than highly specific single component therapies [5-7]. These drug combinations are pharmaco-dynamically synergistic, additive or antagonistic as their effects are greater than, equal to, or less than the summed effects of individual drugs, respectively [8]. These models have garnered interest in the possibility of effective combinatorial drug discovery based on systems-oriented approaches [9-13].

Geva-Zatorsky et al. found that protein responses to combinations of drugs were described accurately by a linear superposition (weighted sum) of their responses to each drug alone [14]. With this in mind, I designed a system for multiple signaling pathways targeting combinatorial drug discovery using gene expression profile. I assumed that if there are two different drugs which regulate two different disease-associated pathways individually, combination of them might be effective unless they affect to each other in unanticipated ways. Based on this model, expression pattern matching methods should be a valuable to quantify the degree of functional similarity among genetic perturbation, disease, and drugs. However, despite current data bases of

mRNA expression profiles, which contain thousands of data points, many of which are available to the public, the number of combinatorial drug discovery approaches based on expression profiles is less than might be expected.

Here I introduce the CDA, for predicting combinatorial drug candidates that target multiple signaling pathways. CDA contains 6,100 expression profiles representing 1,309 molecules which were imported from Connectivity Map [15]. When a user submits “up probe sets” and “down probe sets”, CDA starts hyper-geometric tests for signaling pathway gene set enrichment analysis. Next signaling pathway expression pattern analysis and drug set pattern analysis are performed to measure expression pattern similarity between input signatures and 6,100 expression profiles. These analyses focus on the signaling pathways which are selected in the gene set enrichment analysis (the previous step). CDA then generates lists of single drugs and combinatorial drugs showing similar expression patterns. If user input signatures are disease-related significant probe sets, high negative scoring drugs can be considered candidate drugs for treating individuals whose diseased tissues show opposite gene expression aberrations in signalling pathways as the input cell line.

The results are presented in two different formats: a table view of scores and experimental details, and a network view to visualize relationships between signaling pathway entities and known drugs, proteins and diseases. phExplorer, a graphical data visualization software program, allows users to browse the complex relationships in an interactive and dynamic manner, providing clues to how chemicals work synergistically on certain signaling pathways. To validate the technique, I performed two *in vitro* combinatorial drug discovery studies, on non-small cell lung cancer cells and triple-negative breast cancer (TNBC) cells, and succeeded in each case to find combinatorial drug pairs that exerted synergistic effects in cell culture.

Results

Drug combination suggestion through transcription response module analysis

CDA uses gene expression data in cellular models to pinpoint combinatorial drug pairs that can regulate multiple signaling pathways that potentially synergize to cause disease states, or which through alternate pathways compensate to reduce the efficacy of a drug targeting only one pathway. The combinatorial drug possibility is predicted by gene expression pattern comparison within the selected disease-related signaling pathways. The possibility is scored using Kolmogorov-Smirnov statistics. CDA is composed of four steps; 1) Preparing input signatures and gene set enrichment analysis of signaling pathways 2) Pathway expression pattern analysis 3) Drug set pattern analysis 4) Counting of the number of pathways which show positive/negative correlations with input signatures for drug ranking (See methods for more details and Figure II-1). To validate the technique, I performed one *in silico* single drug discovery study and two *in vitro* combinatorial drug discovery studies. In an *in silico* validation for single drug analysis, CDA successfully identified a molecule having similar function (Case one). As

the discovered combinatorial drug pairs were mostly novel, I carried out *in vitro* validations on non-small cell lung cancer cells and triple-negative breast cancer (TNBC) cells (Case two and three).

Case one: Rapamycin for GC-resistance in acute lymphoblastic leukemia (ALL) cells

Wei et al. demonstrated that rapamycin could reverse the glucocorticoid resistance state to sensitive state [16]. To compare the performances of CMap and CDA, I extracted gene expression signatures of glucocordicoid (GC) sensitivity/resistance in acute lymphoblastic leukemia (ALL) cells (thirteen sensitive and sixteen resistance, GDS2493) using two different methods with different p values. As shown in Table II-1, CMap is highly dependent on extraction method. With signatures using signal-to-noise statistics which is the method Wei et al. used, Rapamycin ranked on second position. However, the ranking fell to 146th and 307th with signatures from Limma, with p value < 0.01 and p value < 0.05, respectively. On the other hand, CDA shows more stable performance. Rapamycin was ranked in top 10 with any of signatures in CDA. This is because unlike CMap consider gene signatures as a set, CDA selectively choose genes

participate on signaling pathways, and treat them as signaling pathway gene sets.

Although genomewide expression analysis with DNA microarray has become a routine tool in genomic research, extracting biologically meaningful information remains a major challenge. Statistically significant genes can be obtained by number of different ways. And there is no standard rule to restrict the number of genes. Significant gene selection is quite depending on individual researchers. As there are multiple ways, significant gene lists are diverse according to extraction algorithms and research principles. This diversity has the risk of insufficient information usage and could lead to inaccurate final interpretation. So I hypothesized that it is more appropriate to use functionally important genes rather than entire statistically selected genes for expression analysis and interpretation. And this hypothesis was validated by this rapamycin case.

Case two: Molecules function as estrogen antagonist

Elevated blood levels of estrogen is associated with an increased risk of breast cancer [17]. Gene expression signatures in breast cancer cells treated with Letrozole (fifty eight untreated tumors and fifty eight letrozole-treated tumors, GDS3116) were used to search the molecules function as estrogen antagonist [18]. Letrozole inhibits, aromatase, an

enzyme that participates in estrogen biosynthesis. By inhibiting estrogen synthesis, letrozole slows the proliferations of breast cancer cells. Table II-2 shows that cells treated with fulvestrant share a very similar expression pattern to those treated with letrozole. Fulvestrant is an estrogen receptor antagonist with no agonist effects. Fulvestrant not only down-regulates transcriptional activities of estrogen receptor but also induce its degradation. Fulvestrant was approved by the FDA for the treatment of postmenopausal women with hormone receptor-positive metastatic breast cancer [19]. The gene expression signatures of cells treated with fulvestrant in 6 different signaling pathways resembled those of letrozole. Not surprisingly, they show similar patterns of gene expression on the plasma membrane estrogen receptor signaling pathway as well as on LPA receptor mediated events pathway and stabilization, expansion of the E-cadherin adherens junction pathway, and Reelin signaling pathway.

These results are illuminating in light of the connections in the literature which show these pathways are regulated by estrogen and/or involved in cancer progression. E-cadherin is a cell-cell adhesion protein, and has been shown to play a crucial role in tumor suppression [20]. A recent study by Oesterrich et al. showed that estrogen caused down-regulation of E-cadherin levels in breast cancer cells [21].

Lysophosphatidic acid (LPA; 1-acyl-glycerol 3-phosphate), which is also regulated by estrogen [22,23] is one of the simplest natural phospholipids that mediates multiple processes including neurogenesis, angiogenesis, wound healing, and cancer progression [24,25]. Reelin is a secreted signaling protein associated with regulation of neuronal cell positioning and migration. Its down-regulation is associated with increased migratory ability and reduced survival in breast cancer [26]. The relationship between reelin and estrogen/breast cancer is not fully understood.

Letrozole inhibits estrogen synthesis, whereas fulvestrant blocks the estrogen receptor. Although the mechanisms of those two compounds are different, the signaling cascades they affect would be expected to be similar in their down-regulation of transcriptional activity in downstream pathways. As levels of estrogen are decreased after the treatment of letrozole, signaling pathways related to E-cadherin and LPA are affected, and this perturbation in these pathways are also observed in cells treated with fulvestrant. They both regulate reelin signaling pathway to induce apoptosis in cancer cells through as yet unknown mechanisms.

Case three: Combinatorial drugs that induce apoptosis on tumorigenic

lung cancer cells

This case derived from a study by Landi et al. that investigated the role of cigarette smoking in lung adenocarcinoma development and survival (forty nine normal lung tissues and fifty eight lung tumor tissues, GDS3257). In our analysis, we disregarded information on smoking, disease state, and gender of the patients. In order to identify molecules that could reverse the expression pattern of lung adenocarcinoma cells, we looked for a phenotype where expression of signature genes was reversed: up-regulated genes became down-regulated, and vice versa. Signaling pathway gene set enrichment analysis of the "reversed phenotype" genes in the lung adenocarcinoma cells showed highlighted that many of the genes identified in this way are frequently associated with tumor cell growth and proliferation (Table II-3). Based on this gene expression analysis, we identified, among the top 15 combinatorial drug pair candidates, two synergistic combinatorial drug pairs: alsterpaullone and scriptaid; and irinotecan and semustin. Alsterpaullone is a cyclin-dependent kinase (CDK) inhibitor that induces apoptosis [27]. Scriptaid is a class of histone deacetylase inhibitors (HDACis). HDACis are involved in cell growth, apoptosis and differentiation. Scriptaid also induces cell death in cancer cells [28,29]. Irinotecan is an anticancer drug that binds to the DNA

topoisomerase 1 complex during DNA replication, preventing the resealing of single-strand breaks [30]. Semustine also known as methyl-CCNU, is another anti-cancer drug in the class of alkylating agents [31,32]. The alsterpaullone-scriptaid and irinotecan-semustine pairs showed meaningful, statistically significant expression pattern matching in seven, six lung adenocarcinoma-related pathways, respectively. Simultaneous and continuous exposure of A549 cells to different concentration of these two combinatorial drug pairs for 72 hours showed a synergism (Combination index (CI) < 1 and Dose reduction index (DRI) > 1; Table II-4 and II-5, Figure II-2).

Case four: Combinatorial drugs that induce apoptosis on triple-negative breast cancer cells

Breast cancer is the most common form of cancer in women. Human epidermal growth factor receptor 2 (HER2), also known as receptor tyrosine-protein kinase ERBB2, belongs to the epidermal growth factor receptor (EGFR) family, and it is one of the most important oncogenes in invasive breast cancer. Based on the importance of HER2 amplification on breast cancer, the HER2-targeting monoclonal antibody trastuzumab was developed [33]. Additionally, aberrant EGFR signaling is a major characteristic of a

human cancer including breast cancer. Several anti-EGFR agents are currently undergoing clinical testing in breast cancer patients clinically [34]. However, triple negative breast cancer (TNBC) is a type of breast cancers that does not express the genes for estrogen receptor (ER), progesterone receptor (PR) or human epidermal growth factor receptor 2 (HER2). For that reason, novel effective therapeutic agents are needed for TNBC patients [35]. Combined treatment of general breast cancer cells with drugs that target EGFR and HER2 results in a synergistic antitumor effect [36,37]. That means that targeting EGFR family signaling pathway is a good strategy for breast cancer treatment.

To discover a synergistic combinatorial drug pair for TNBC patients, I focused on FDA approved drugs. I obtained gene expression signatures from TNBC cell lines (five normal breast cancer cell lines and five triple-negative breast cancer cell lines, GSE6569), and halofantrine - vinblastine pair were selected as a candidate pair (Figure II-3). The CDA analysis indicated that the pair has opposite expression patterns compared with TNBC signatures in five different signaling pathways, including four of the EGFR family signaling pathways and one integrin pathway (Figure II-4). Aberrant activation of the EGFR family is implicated in a number of cancers and it is already the

target of several antineoplastic agents [38]. A6b1- and a6b4- mediated integrin signaling is involved in apoptosis, tumour cell invasions, and cell migration.

Halofantrine is an anti-malarial agent with an unknown mode of action. Although it has cardiotoxic potential, it is safe when carefully administered [39]. Vinblastine is a microtubule-targeted anticancer drug that induces mitotic block and apoptosis by suppressing microtubule dynamics at lower concentration, and reducing microtubule polymer mass at higher concentration [40]. As shown in Figure II-4B, halofantrine and vinblastine are indirectly related to EGFR family signaling pathways. Furthermore, both are also related to an integrin signaling pathway. Based on this information, I hypothesized that halofantrine and vinblastine are synergistic because they simultaneously affect the EGFR and integrin signaling pathways. Furthermore, sensitivity of HER2-positive breast cancer cells resistant to anti-HER2 therapies are related to antiapoptotic proteins MCL1 and Survivin [41]. And these two proteins commonly have protein-protein interactions with CASP3, a vinblastine-related protein [42,43]. Based on this, I hypothesized that vinblastin could be a good TNBC drug candidate. Using the steps described for all three cases, CDA users will be able to put forward testable hypotheses by combining signaling pathway expression information

with known drug-protein-disease information from phExplorer.

Figure II-1. Analysis pipeline of CDA

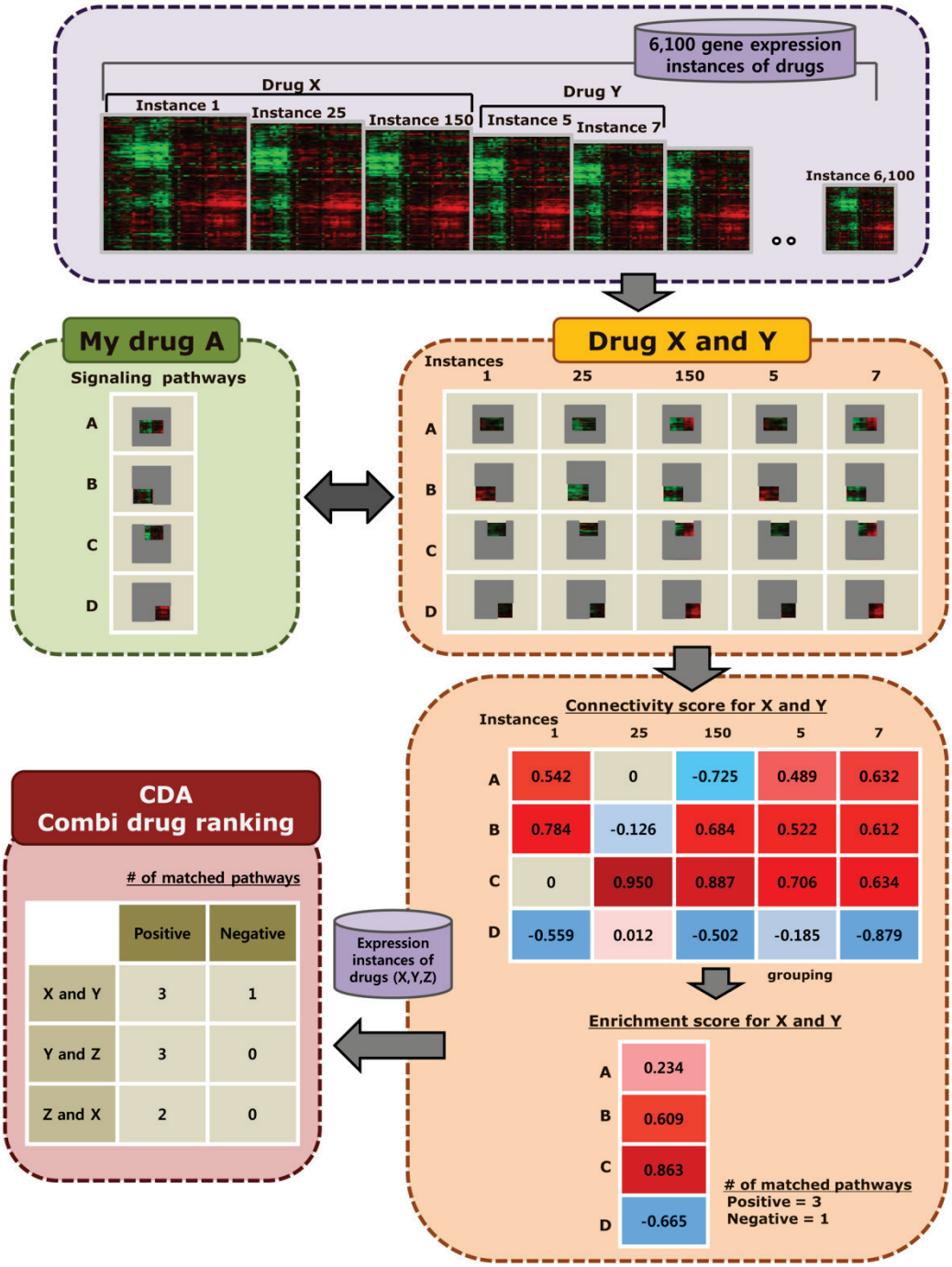


Figure II-1. Analysis pipeline of CDA. Combinatorial drug analysis process. In drug set pattern analysis step (the bottom right box), combinatorial drug analysis process treats profiles of two different molecules as a group to measure the synergistic effects of them.

Figure II-2. Synergistic combinatorial drug pairs on lung cancer cells

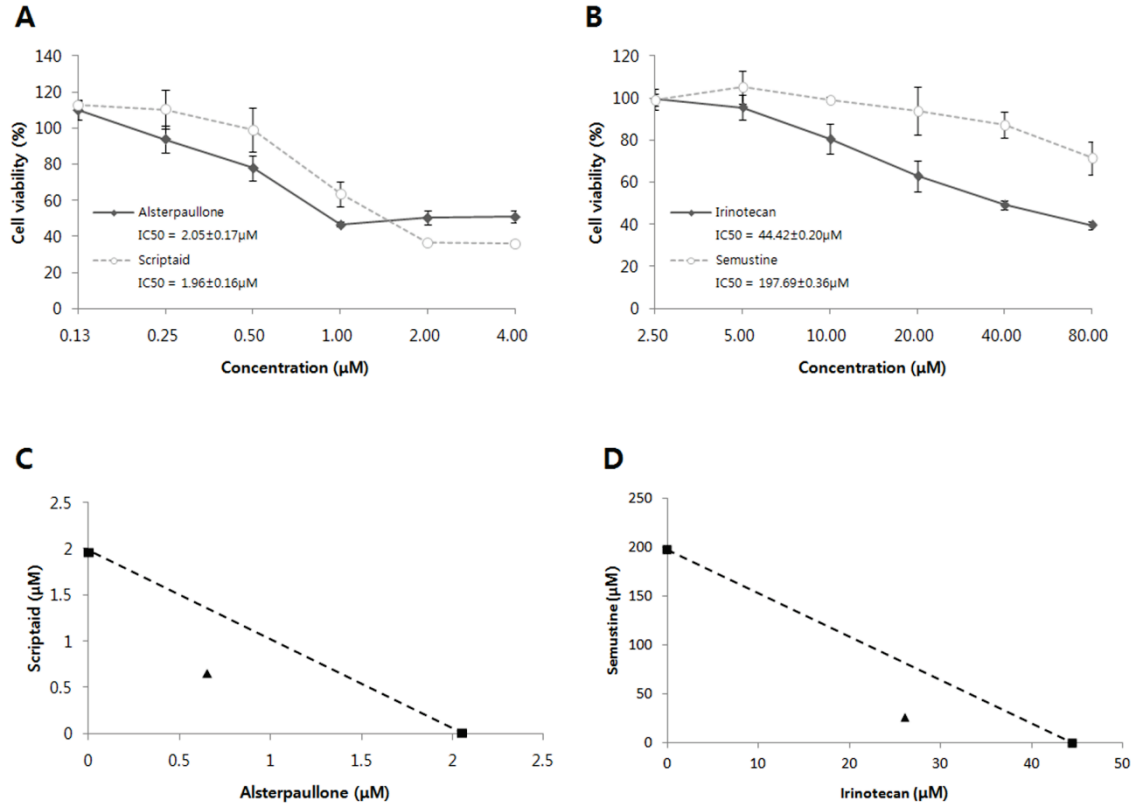


Figure II-2. Synergistic combinatorial drug pairs on lung cancer cells. (A, B) Effects of alsterpaullone, scriptaid, irinotecan, and semustine on A549 cancer cell proliferation. IC50 indicates the concentration of drug that induce 50% of inhibition of cell proliferation. Error bars represent the standard deviation of six experiments.

(C, D) Drug pairs were treated in 1:1 molar ratio. The IC50 values of each drug are plotted on the axes, and the dashed line represents addictive effect. Triangle point represents the concentrations of the combinations resulting in 50% of proliferation inhibition. As the triangle points are positioned on the left of the dashed line, these combinatorial drug pairs are synergistic. The IC50 values of each drug in alsterpaullone-scriptaid and irinotecan-semustine combinations are 0.65μM and 26.05μM, respectively.

Figure II-3. In vitro validation of halofantrine and vinblastine alone and in combination in a triple-negative breast cancer cell line

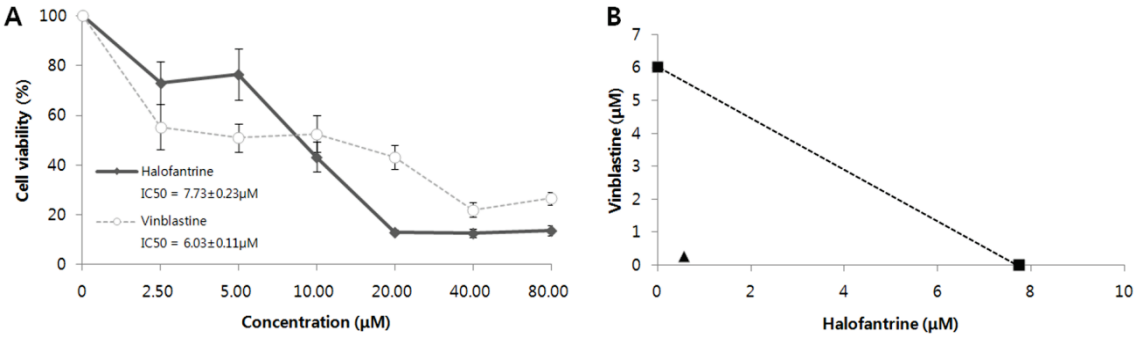


Figure II-3. In vitro validation of halofantrine and vinblastine alone and in combination in a triple-negative breast cancer cell line. (A) Effects of halofantrine and vinblastine on MDA-MB-231 TNBC cell proliferation. IC50 indicates the concentration of drug that induce 50% of inhibition of cell proliferation.

(B) Halofantrine and vinblastine combination was treated in 2:1 molar ratio. Halofantrine and vinblastine combination shows a strong synergistic effect. The IC50 values of each drug in halofantrine-vinblastine combinations are 0.55μM and 0.27μM, respectively. The combination shows a strong synergistic effect (CI value is 0.12, and DRI values for halofantrine and vinblastine are 14.17 and 22.09, respectively).

Figure II-4. Network map of halofantrine and vinblastine on triple-negative breast cancer using phExplorer

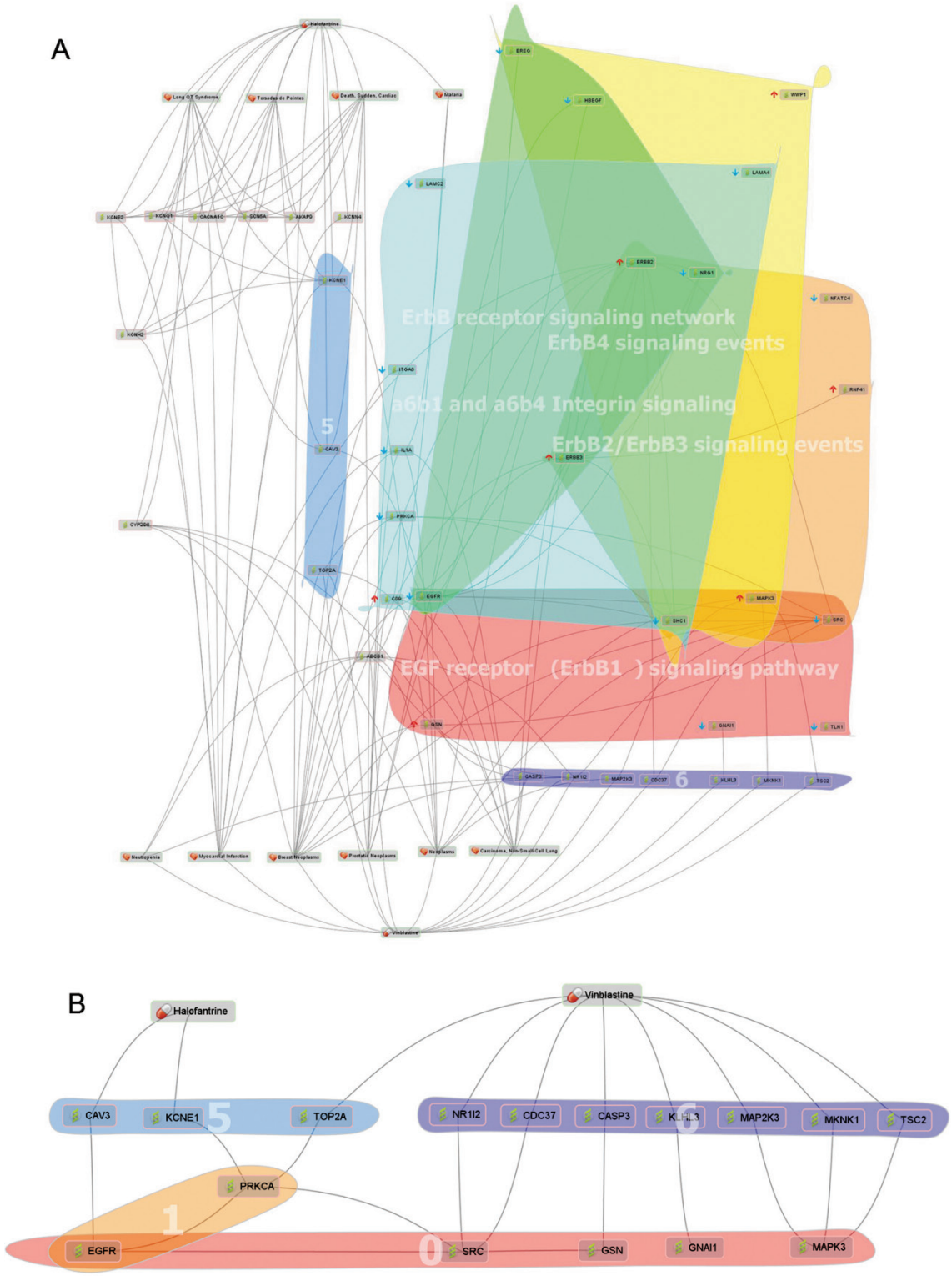


Figure II-4. Network map of halofantrine and vinblastine on triple-negative breast cancer using phExplorer. (A) It seems that halofantrine and vinblastine could affect on five different signaling pathways in TNBC. Group 5: Halofantrine- or vinblatine-related proteins which are also related with proteins of A6B1 and A6B4 Integrin signaling pathway. Group 6: Proteins which are related with vinblasitne as well as proteins of EGFR family signaling pathways (such as ERBB1 signaling pathway, ERBB2/ERBB3 signaling events, ERBB4 signaling events, ERBB receptor signaling network).

(B) We hypnotized that halofantrine and vinblastine are synergistic because they complementary regulate integrin and EGFR signaling pathways. Group 0: A part of EGFR family signaling pathways. Group 1: A part of A6B1 and A6B4 Integrin signaling pathway.

Table II-1. Ranking of rapamycin in GC-resistant ALL cells

DEG extraction method	Num of signatures	Rank in CMap	Rank in CDA
Signal-to-noise (p <= 0.0005)	157	2	Not found
Signal-to-noise (p <= 0.001)	244	2	10
Limma (p <= 0.01)	391	146	5
Limma (p <= 0.05)	543	307	6

Table II-2. Top 10 molecules showing similar expression patterns of transcriptional response modules to letrozole

	FOXM1 transcription factor network	IGF1 pathway	IL4-mediated signaling events	LPA receptor mediated events	Plasma membrane estrogen receptor signaling	Reelin signaling pathway	Stabilization and expansion of the E-cadherin adherens junction
Fulvestrant		O	O	O	O	O	O
Trichostatin A	O	O	O		O	O	
Irinotecan	O	O		O	O		O
Ag-013608					O	O	O
Tretinoin	O	O			O		
Metamizole sodium		O			O		O
Tanespimycin		O	O	O			
Vorinostat	O	O			O		
Verteporfin		O			O		
Daunorubicin		O					O

Table II-3. Enriched pathway in lung adenocarcinoma

Pathway	Pathway Category
	Integrin mediated cell-cell signaling pathways
amb2 Integrin signaling	Integrin mediated cell-extracellular matrix signaling pathways
Aurora A signaling	Cell cycle pathways, mitotic
Aurora B signaling	Cell cycle pathways, mitotic
BMP receptor signaling	Bone morphogenetic proteins signaling pathway
Direct p53 effectors	p53 signaling pathway
	Transcription factor mediated signaling pathways
E2F transcription factor network	Cell cycle pathways, mitotic
	Transcription pathways
Endothelins	Endothelin signaling pathway
FGF signaling pathway	Fibroblast growth factor signaling pathway
FOXM1 transcription factor network	Forkhead signaling pathways

Table II-4. CI values for the drug combinations at 25%, 50%, 75% levels of inhibition of A549 cell proliferation

CI Values	25%	50%	75%
Alsterpaullone + Scriptaid	0.887	0.647	0.483
Irinotecan + Semustine	0.816	0.718	0.636

Table II-5. DRI values for the drug combinations at 25%, 50%, 75% levels of inhibition of A549 cell proliferation

DRI Values	25%	50%	75%
Alsterpaullone + Scriptaid			
Alsterpaullone	2.013	3.162	4.968
Scriptaid	2.565	3.020	3.554
Irinotecan + Semustine			
Irinotecan	1.452	1.705	2.002
Semustine	7.841	7.589	7.345

Discussion

Since the number of new drug has not kept pace with the enormous increase in pharma R&D spending, drug discovery researchers have become more creative in finding new uses for existing drugs [44]. Analyzing large data sets such as gene expression [15], chemical similarity [45], side-effect similarity [46], disease-drug network [47], and phenotypic disease network [48] has been applied for drug repositioning. Exploration of drug off-targets using chemical-protein interactome can also provide alternative strategy [49]. However drugs with single targets frequently show limited efficacies and drug resistance at the some point. To overcome these problems, systems-oriented drug design is now moving to multicomponent therapies and multi-targeted drugs, based on the idea that targeting drugs to act on multiple signaling pathways will maximize therapeutic efficacy [50]. With this in mind, I have designed a system for multiple signaling pathways targeting combinatorial drug discovery using gene expression profile. There are three groups of pharmacodynamically synergistic combinations; 1) anti-counteractive action group 2) complementary action group 3) facilitating action group. There are a variety of mechanism of actions represented by these combinations, arising

from drug interactions with the same or different targets of the same or different pathways, and from modulations of crosstalk pathways and network robustness [8].

The robustness of CDA does not depend heavily on the particular bioinformatics method employed for signature extraction, thus providing a flexible analysis platform that can be adopted by a variety of users with different software tools for handling gene expression analysis. Although genome-wide expression analysis has become a routine tool in genomic research, extracting biologically meaningful information remains a major challenge. Statistically significant genes can be obtained by number of different ways. Moreover, there is no standard rule to restrict the number of genes. Thus, significant gene selection is quite depending on individual researchers. Given this multiplicity of approaches, significant gene lists can be quite diverse according to extraction algorithms and research principles. This lack of standardized bioinformatics approaches brings with it a risk of insufficient information usage that can lead to inaccuracies in the final interpretation. To offset these differences, for expression analysis and interpretation, our strategy employs functionally important genes as data sets, rather than entire statistically selected gene sets. This approach was validated by an *in silico* case (Case one). CDA provides a mechanism whereby hundreds of input

signature genes will be split into signaling pathways at the first step, therefore users don't need to themselves extract a small group of significant gene sets using number of different algorithms. Through this process, CDA successfully has identified a number of molecules having similar function (Table II-2). In this study, I presented case studies whereby CDA successfully predicted synergistic combinatorial drug pairs in lung cancer and triple negative breast cancer. Together with phExplorer, CDA also provides functional insights of combinatorial drugs.

Using CDA, the number of matched pathways decides the ranking of drug candidates, however, the type of matched pathways must be considered carefully. As the interpretation of result and the final decision must be made by researchers, I tried not to restrict their choice by providing strictly ordered list based on our limited pre-knowledge.

Materials and Methods

Data source

Reference molecule-treated expression data was downloaded from Connectivity Map (build 02) (http://www.broadinstitute.org/cmap/). It contains 6,100 expression profiles representing 1,309 molecules. Molecules were selectively applied to five different human cancer cell lines for short duration. Each molecule-treated expression profile was paired with a control, and each profile was represented by a non-parametric rank-ordered list of all probe sets.

Pathway gene set data was downloaded from Pathway Interaction Database (PID) on 09/03/2010 (http://pid.nci.nih.gov/). Only the NCI-Nature Curated data was used. Pathway gene set information was extracted, consisting of 166 pathways comprising 2,297 genes. These genes were annotated to Affymetrix GeneChip Human Genome U133 Array Set HG-U133A probe set. The final form of pathway data consists of 166 signaling pathways and 3,726 probe sets.

Furthermore, nine public databases, EntrezGene interaction[51], MINT[52], DIP[53], CTD[54], TTD[55], ChemBank[56], PharmGKB[57], OMIM

(http://www.ncbi.nlm.nih.gov/omim/), and GAD[58] were integrated to visualise enrich drug-protein-disease network map. For data integration in a unified format, we adopted PubChem CID for drugs, GeneID for proteins, and MeSH descriptor for diseases. The integrated database is called PharmDB, and it is available at http://pharmdb.org/.

Input signatures

Three different GDS/GSE data files were downloaded for each case study. All of them were used Affymetrix Human Genome U133A Array.

Case 1: GDS3116 - Letrozole effect on breast cancer

Fifty eight untreated tumors vs. fifty eight letrozole-treated tumors

Case 2: GDS3257 - Lung adenocarcinoma

Forty nine normal lung tissues vs. fifty eight lung tumor tissues

Case 3: GSE6569 - Triple-negative breast cancer cell lines

Five normal breast cancer cell lines: BT474, SKBR3, HCC-1419, HCC-1954, MCF7

Triple-negative breast cancer cell lines: BT20, BT549, HCC-1806, MDA-MB-231, MDA-MB-468

The expression data were normalized using RMA from the BioConductor Affy package. Then these data were analyzed using a method called empirical Bayes in limma. To extract statistically differentially expressed genes, 2-fold change and p-value < 0.05 were set as default. The signatures were represented by two probe sets, “up probe sets” and “down probe sets”. With given input signatures, hyper geometric tests were performed for signaling pathway gene set enrichment analysis. Signaling pathways with p-value < 0.01 were selected as it was believed that input signature genes were enriched in these pathways.

Enrichment Analysis

Signaling pathway expression pattern analysis and drug set pattern analysis were performed based on the Kolmogorov-Smirnov statistics. To determine whether the distribution of input gene sets/or drug sets was significant, 10,000 times permutations were carried out by generating random ranking matrices. The sets with p-value < 0.01 were indicated as enriched.

Signaling pathway expression pattern analysis

6,100 molecule-treated expression profiles were rank ordered using gene set enrichment analysis for each selected pathway. As mentioned above, there were two types of input set, “up probe sets” and “down probe sets”. The expression pattern similarity is calculated for both sets. The procedure is as follows:

1) Calculate Kolmogorov-Smirnov score for both “up probe sets” and “down probe sets”

e = an expression profile

KS^e = KS (Kolmogorov-Smirnov) score for the “up probe sets” or the “down probe sets”

n = the total number of probe sets (22,283)

t = the number of probe sets in either the “up probe sets” or the “down probe sets”

j = the position of a probe set in the ordered input signature probe set lists

V(j) = the position of the jth probe set in the ordered list of all probe sets.

$$a = \underset{j=1}{\overset{t}{Max}}\left[\frac{j}{t} - \frac{V(j)}{n}\right]$$

$$b = \underset{j=1}{\overset{t}{Max}}\left[\frac{V(j)}{n} - \frac{(j-1)}{t}\right]$$

$$KS^e = \begin{cases} a\,(\text{if } a > b) \\ -b\,(\text{if } b > a) \end{cases}$$

2) Calculate the Enrichment Score (ES) for each profile

$$ES^e = 0\ (if\ KS_{up}\ and\ KS_{down}\ \text{have the same algebraic sign})$$

Otherwise, across all profiles,

$$s^e = KS_{up} - KS_{down}$$
$$p = Max(s^e)$$
$$q = Max(s^e)$$

The ES for these profiles are:

$$ES^e = \left\{ \begin{array}{l} \frac{s^e}{p} (\text{if } s^e > 0) \\ -\left(\frac{s^e}{q}\right) (\text{if } s^e < 0) \end{array} \right\}$$

3) Rank the profiles in descending order of ES^e

Drug set pattern analysis

Molecules were applied to different cell lines with various doses, and the ES of each molecule was calculated using the distribution of the molecule-treated profiles, using the same method as used in calculating the KS score in signaling pathway expression pattern comparison. For the case of combinatorial drug analysis, signatures of two different molecules were treated as a group. The rationale is as follows: we assume two molecules, “A” and “B” show highly similar expression pattern with the expression of signaling pathway “SP1” and “SP2”, respectively. The purpose of combinatorial drug

is matching up two molecules which are synergistic or complementary. “A” and “B” are highly related with different pathways, and thus might affect to each other in unanticipated ways. For that reason, profiles of “A” and “B” are grouped as a set, then the ES (Enrichment Score) of “A and B” combination is calculated in two signaling pathways independently. So the similarity of expression pattern of “B” is now considered not only in “SP2” but also in “SP1” as a combinatorial drug partner. If “B” shows high ESs in both pathways, “B” could be a complementary partner for “A” as it covers “SP2” which “A” might not be able to regulate, and at the same time, synergistic effect could be expected in “SP1” as both of them are highly enriched in there.

Using these steps, the KS score was computed using these profiles. Then, random permutation tests (10,000 times) were carried out to estimate the significance of a distribution of those profiles. The molecules with p-value < 0.01 were assumed as significant.

Drug ranking

At this point, I have listed single/combinatorial drugs for each disease-associated signaling pathway in our database. The goal of creating this system is to provide a

means of selecting single/combinatorial drugs that can regulate disease-related signaling pathways to the greatest potential. To this end, for each drug, the number of pathways scored greater than the positive threshold was counted. The positive threshold for single drug and combinatorial drug were 0 and 0.5, respectively. The drugs were ranked in descending order of the number of pathways they appeared in. Pathways that scored less than the negative threshold were also listed. The negative threshold for single drug and combinatorial drug were 0 and -0.5, respectively. These negatively correlated pathways can be treated as negative effects.

Cell Culture and Materials

A549 and MDA-MB-231 were purchased from American Type Culture Collection. RPMI containing 10% fetal bovine serum and 1% antibiotics were used for cell cultivation. Alsterpaullone, Scriptaid, Irinotecan hydrochloride, Semustine, Halofantrine hydrochloride, Vinblastine sulfate salt were purchased from Sigma.

MTT Assay

A549 or MDA-MB-231 cells were seeded in the 96-well plates. After 24 h, cells were

treated with indicated chemicals. After incubation for 3 days, MTT reagent (5 mg/ml) (Sigma) was added to each well, and the plate was placed at 37oC for 2 h. After aspirating the supernatant, 200μl of dimethyl sulfoxide (Sigma) was added to each well. Colored formazan product was assayed spectrophotometrically at 570 nm using ELISA plate reader.

Combination index (CI) and Dose reduction index (DRI) calculations

Synergism and antagonism for combinatorial drug were quantified by the combination index (CI), where CI < 1, CI = 0, CI > 0 indicate synergism, addictive, and antagonism, respectively. CI was determined by the following equation:

$$CI_{A+B} = \frac{D_{A/A+B}}{D_A} + \frac{D_{B/A+B}}{D_B}$$

D_A is the concentration of drug A that induce the inhibition of cell growth. $D_{A/A+B}$ is the concentration of drug A in the combination A+B giving the same inhibition effect. The dose reduction index (DRI) is a measure of how much the dose of each drug may be reduced in a combination for a given degree of effect compared with the concentration of each drug alone.

$$DRI_A = \frac{D_A}{D_{A/A+B}} \qquad \text{and} \qquad DRI_B = \frac{D_B}{D_{B/A+B}}$$

The CI and DRI indexes were calculated with the CalcuSyn version 2.1 software (Biosoft, Cambridge, UK).

Acknowledgments

This study was supported by the grants of the Global Frontier (NRF-M1AXA002-2010-0029785) and the Research Information Center Supporting Program (2012-0000350) and the WCU project (R31-2008-000-10103-0) of the Ministry of Education, Science, and Technology and Korea Healthcare Technology (A092255-0911-1110100), the Ministry of Health and Welfare Affairs, and Gyonggi-do to SK, an EU project of the 7th framework programme (METOXIA) to CTS, and by the Korean Ministry of Education, Science and Technology (MEST) under grant number 20110002321.

References

1. Kubinyi H: **Drug research: myths, hype and reality**. *Nat Rev Drug Discov* 2003, **2**(8):665-668.

2. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MW, Silver RT, Goldman JM, Stone RM *et al*: **Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia**. *N Engl J Med* 2006, **355**(23):2408-2417.

3. Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J: **Robustness of cellular functions**. *Cell* 2004, **118**(6):675-685.

4. Kitano H: **A robustness-based approach to systems-oriented drug design**. *Nat Rev Drug Discov* 2007, **6**(3):202-210.

5. Gupta EK, Ito MK: **Lovastatin and extended-release niacin combination product: the first drug combination for the management of hyperlipidemia**. *Heart Dis* 2002, **4**(2):124-137.

6. Larder BA, Kemp SD, Harrigan PR: **Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy**. *Science* 1995,

269(5224):696-699.

7. Nelson HS: **Advair: combination treatment with fluticasone propionate/salmeterol in the treatment of asthma**. *J Allergy Clin Immunol* 2001, **107**(2):398-416.

8. Jia J, Zhu F, Ma X, Cao Z, Li Y, Chen YZ: **Mechanisms of drug combinations: interaction and network perspectives**. *Nat Rev Drug Discov* 2009, **8**(2):111-128.

9. Hahn CK, Ross KN, Warrington IM, Mazitschek R, Kanegai CM, Wright RD, Kung AL, Golub TR, Stegmaier K: **Expression-based screening identifies the combination of histone deacetylase inhibitors and retinoids for neuroblastoma differentiation**. *Proc Natl Acad Sci U S A* 2008, **105**(28):9751-9756.

10. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, Rosen N, Gennemark P, Sander C: **Models from experiments: combinatorial drug perturbations of cancer cells**. *Mol Syst Biol* 2008, **4**:216.

11. Chatterjee MS, Purvis JE, Brass LF, Diamond SL: **Pairwise agonist scanning predicts cellular signaling responses to combinatorial stimuli**. *Nat Biotechnol* 2010, **28**(7):727-732.

12. Zhao XM, Iskar M, Zeller G, Kuhn M, van Noort V, Bork P: **Prediction of drug**

combinations by integrating molecular and pharmacological data. *PLoS Comput Biol* 2011, **7**(12):e1002323.

13. Wu Z, Zhao XM, Chen L: **A systems biology approach to identify effective cocktail drugs**. *BMC Syst Biol* 2010, **4 Suppl 2**:S7.

14. Geva-Zatorsky N, Dekel E, Cohen AA, Danon T, Cohen L, Alon U: **Protein dynamics in drug combinations: a linear superposition of individual-drug responses**. *Cell* 2010, **140**(5):643-651.

15. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN *et al*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**(5795):1929-1935.

16. Wei G, Twomey D, Lamb J, Schlis K, Agarwal J, Stam RW, Opferman JT, Sallan SE, den Boer ML, Pieters R *et al*: **Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance**. *Cancer Cell* 2006, **10**(4):331-342.

17. Clemons M, Goss P: **Estrogen and the risk of breast cancer**. *N Engl J Med* 2001, **344**(4):276-285.

18. Miller WR, Larionov AA, Renshaw L, Anderson TJ, White S, Murray J, Murray E, Hampton G, Walker JR, Ho S *et al*: **Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole**. *Pharmacogenet Genomics* 2007, **17**(10):813-826.

19. Croxtall JD, McKeage K: **Fulvestrant: a review of its use in the management of hormone receptor-positive metastatic breast cancer in postmenopausal women**. *Drugs* 2011, **71**(3):363-380.

20. Berx G, Van Roy F: **The E-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression**. *Breast Cancer Res* 2001, **3**(5):289-293.

21. Oesterreich S, Deng W, Jiang S, Cui X, Ivanova M, Schiff R, Kang K, Hadsell DL, Behrens J, Lee AV: **Estrogen-mediated down-regulation of E-cadherin in breast cancer cells**. *Cancer Res* 2003, **63**(17):5203-5208.

22. Hama K, Aoki J, Bandoh K, Inoue A, Endo T, Amano T, Suzuki H, Arai H: **Lysophosphatidic receptor, LPA3, is positively and negatively regulated by progesterone and estrogen in the mouse uterus**. *Life Sci* 2006, **79**(18):1736-1740.

23. Gonzalez-Arenas A, Avendano-Vazquez SE, Cabrera-Wrooman A, Tapia-Carrillo D,

Larrea F, Garcia-Becerra R, Garcia-Sainz JA: **Regulation of LPA receptor function by estrogens**. *Biochim Biophys Acta* 2008, **1783**(2):253-262.

24. Contos JJ, Ishii I, Chun J: **Lysophosphatidic acid receptors**. *Mol Pharmacol* 2000, **58**(6):1188-1196.

25. Moolenaar WH: **Bioactive lysophospholipids and their G protein-coupled receptors**. *Exp Cell Res* 1999, **253**(1):230-238.

26. Stein T, Cosimo E, Yu X, Smith PR, Simon R, Cottrell L, Pringle MA, Bell AK, Lattanzio L, Sauter G *et al*: **Loss of reelin expression in breast cancer is epigenetically controlled and associated with poor prognosis**. *Am J Pathol* 2010, **177**(5):2323-2333.

27. Lahusen T, De Siervi A, Kunick C, Senderowicz AM: **Alsterpaullone, a novel cyclin-dependent kinase inhibitor, induces apoptosis by activation of caspase-9 due to perturbation in mitochondrial membrane potential**. *Mol Carcinog* 2003, **36**(4):183-194.

28. Lee EJ, Lee BB, Kim SJ, Park YD, Park J, Kim DH: **Histone deacetylase inhibitor scriptaid induces cell cycle arrest and epigenetic change in colon cancer cells**. *Int J Oncol* 2008, **33**(4):767-776.

29. Brazelle W, Kreahling JM, Gemmer J, Ma Y, Cress WD, Haura E, Altiok S: **Histone deacetylase inhibitors downregulate checkpoint kinase 1 expression to induce cell death in non-small cell lung cancer cells**. *PLoS One* 2010, **5**(12):e14335.

30. Marsh S, Hoskins JM: **Irinotecan pharmacogenomics**. *Pharmacogenomics* 2010, **11**(7):1003-1010.

31. Guo Y, Lu JJ, Ma X, Wang B, Hong X, Li X, Li J: **Combined chemoradiation for the management of nasal natural killer (NK)/T-cell lymphoma: elucidating the significance of systemic chemotherapy**. *Oral Oncol* 2008, **44**(1):23-30.

32. Zhao Z, Liu Y, He H, Chen X, Chen J, Lu YC: **Candidate genes influencing sensitivity and resistance of human glioblastoma to Semustine**. *Brain Res Bull* 2011, **86**(3-4):189-194.

33. Bange J, Zwick E, Ullrich A: **Molecular targets for breast cancer therapy and prevention**. *Nat Med* 2001, **7**(5):548-552.

34. Lo HW, Hsu SC, Hung MC: **EGFR signaling pathway in breast cancers: from traditional signal transduction to direct nuclear translocalization**. *Breast Cancer Res Treat* 2006, **95**(3):211-218.

35. Gluz O, Liedtke C, Gottschalk N, Pusztai L, Nitz U, Harbeck N: **Triple-negative breast cancer--current status and future directions**. *Ann Oncol* 2009, **20**(12):1913-1927.

36. Normanno N, Campiglio M, De LA, Somenzi G, Maiello M, Ciardiello F, Gianni L, Salomon DS, Menard S: **Cooperative inhibitory effect of ZD1839 (Iressa) in combination with trastuzumab (Herceptin) on human breast cancer cell growth**. *Ann Oncol* 2002, **13**(1):65-72.

37. Moulder SL, Yakes FM, Muthuswamy SK, Bianco R, Simpson JF, Arteaga CL: **Epidermal growth factor receptor (HER1) tyrosine kinase inhibitor ZD1839 (Iressa) inhibits HER2/neu (erbB2)-overexpressing breast cancer cells in vitro and in vivo**. *Cancer Res* 2001, **61**(24):8887-8895.

38. Zhang H, Berezov A, Wang Q, Zhang G, Drebin J, Murali R, Greene MI: **ErbB receptors: from oncogenes to targeted cancer therapies**. *J Clin Invest* 2007, **117**(8):2051-2058.

39. Bouchaud O, Imbert P, Touze JE, Dodoo AN, Danis M, Legros F: **Fatal cardiotoxicity related to halofantrine: a review based on a worldwide safety data base**. *Malar J* 2009, **8**:289.

40. Jordan MA, Wilson L: **Microtubules as a target for anticancer drugs**. *Nat Rev Cancer* 2004, **4**(4):253-265.

41. Valabrega G, Capellero S, Cavalloni G, Zaccarello G, Petrelli A, Migliardi G, Milani A, Peraldo-Neia C, Gammaitoni L, Sapino A *et al*: **HER2-positive breast cancer cells resistant to trastuzumab and lapatinib lose reliance upon HER2 and are sensitive to the multitargeted kinase inhibitor sorafenib**. *Breast Cancer Res Treat* 2011, **130**(1):29-40.

42. Tamm I, Wang Y, Sausville E, Scudiero DA, Vigna N, Oltersdorf T, Reed JC: **IAP-family protein survivin inhibits caspase activity and apoptosis induced by Fas (CD95), Bax, caspases, and anticancer drugs**. *Cancer Res* 1998, **58**(23):5315-5320.

43. Weng C, Li Y, Xu D, Shi Y, Tang H: **Specific cleavage of Mcl-1 by caspase-3 in tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)-induced apoptosis in Jurkat leukemia T cells**. *J Biol Chem* 2005, **280**(11):10491-10500.

44. Ashburn TT, Thor KB: **Drug repositioning: identifying and developing new uses for existing drugs**. *Nat Rev Drug Discov* 2004, **3**(8):673-683.

45. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH,

Kuijer MB, Matos RC, Tran TB *et al*: **Predicting new molecular targets for known drugs**. *Nature* 2009, **462**(7270):175-181.

46. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity**. *Science* 2008, **321**(5886):263-266.

47. Hu G, Agarwal P: **Human disease-drug network based on genomic expression profiles**. *PLoS One* 2009, **4**(8):e6536.

48. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA: **A dynamic network approach for the study of human phenotypes**. *PLoS Comput Biol* 2009, **5**(4):e1000353.

49. Yang L, Wang K, Chen J, Jegga AG, Luo H, Shi L, Wan C, Guo X, Qin S, He G *et al*: **Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome--clozapine-induced agranulocytosis as a case study**. *PLoS Comput Biol* 2011, **7**(3):e1002016.

50. Smalley KS, Haass NK, Brafford PA, Lioni M, Flaherty KT, Herlyn M: **Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases**. *Molecular cancer therapeutics* 2006, **5**(5):1136-1144.

51. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2011, **39**(Database issue):D52-57.

52. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update**. *Nucleic Acids Res* 2010, **38**(Database issue):D532-539.

53. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.

54. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ: **The Comparative Toxicogenomics Database: update 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):D1067-1072.

55. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C *et al*: **Update of TTD: Therapeutic Target Database**. *Nucleic Acids Res* 2010, **38**(Database issue):D787-791.

56. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M *et al*: **ChemBank: a small-molecule screening and cheminformatics resource database**. *Nucleic Acids Res* 2008,

36(Database issue):D351-359.

57. Thorn CF, Klein TE, Altman RB: **Pharmacogenomics and bioinformatics: PharmGKB**. *Pharmacogenomics* 2010, **11**(4):501-505.

58. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nat Genet* 2004, **36**(5):431-432.

국문초록

신약재창출을 위한 계산 방법

신약 발굴 과정은 오랜 연구개발 시간과 고비용을 요구하고 있으며, 그 성공 가능성 또한 대단히 낮다. 그러므로 현재 시판되고 있는 기존 약물들의 새로운 적응증을 발굴하는 것에 큰 기대가 쏠리고 있고 이는 신약재창출이라고 부른다. 만약 이러한 신약재창출이 성공한다면, 이로 인해 실패 위험 감소와 함께 초기 개발 비용 절감이라는 두 가지 효과를 기대할 수 있게 된다. 그리고 이러한 과정을 논리적인 방법으로 해결할 수 있는 새로운 방법론 개발이 시급해진 실정이다.

본 연구에서는 PharmDB 라는 질병, 약물, 그리고 단백질 간의 상호 연결 정보를 포함하고 있는 통합형 생명의약학 데이터베이스를 개발하였다. 또한 이러한 정보들의 네트워크 분석을 통해 약물의 새로운 적응증을 예측할 수 있는 Shared Neighborhood Scoring(SNS) 알고리즘을 개발하였다.

더 나아가 최적의 조합약물을 발굴할 수 있는 Combinatorial Drug Assembler (CDA)라고 하는 시스템을 개발하였다. 이는 유전자 발현 정보를 이용하여 multiple signaling pathways 를 타겟팅하는 조합 약물을 발굴하는 시스템이다. 하나의 signaling transduction pathway 를 타겟팅하는 항암치료는 서로 다른 signaling pathways 간의 overlapping 기능이나 cross-talk 때문에 종종

치료에 실패한다. 특정 경우에선 강력하게 하나의 component 를 타겟팅하는 치료보다 balanced multi-component 치료가 더 효과적이라는 사실이 최근 연구를 통해 밝혀졌다. CDA 는 multiple signaling pathways 를 타겟팅하는 최적의 조합약물을 시스템적인 방법으로 발굴하는 새로운 방법론을 제시하였다.

PharmDB 와 CDA 는 각각 http://pharmdb.org/, http://cda.i-pharm.org 에서 현재 서비스 중이다.

주요어: 신약재창출, tripartite network, Shared Neighborhood Scoring (SNS) algorithm, 시스템 생물학, 조합약물, 신호전달 경로, 유전자 발현정보

학번: 2008-31001