BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 10 2014, pages 1456–1463
doi:10.1093/bioinformatics/btu046

Systems biology

Advance Access publication January 24, 2014

Combinatorial therapy discovery using mixed integer linear programming

Kaifang Pang[1,2,3], Ying-Wooi Wan[1,2,4], William T. Choi[1,2], Lawrence A. Donehower[5], Jingchun Sun[6], Dhruv Pant[7] and Zhandong Liu[1,2,3,*]

[1]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, [2]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, [3]Department of Pediatrics-Neurology, [4]Department of Obstetrics and Gynaecology, [5]Department of Molecular Virology and Microbiology, Baylor College of Medicine, [6]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, and [7]Department of Cancer Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Combinatorial therapies play increasingly important roles in combating complex diseases. Owing to the huge cost associated with experimental methods in identifying optimal drug combinations, computational approaches can provide a guide to limit the search space and reduce cost. However, few computational approaches have been developed for this purpose, and thus there is a great need of new algorithms for drug combination prediction.

Results: Here we proposed to formulate the optimal combinatorial therapy problem into two complementary mathematical algorithms, Balanced Target Set Cover (BTSC) and Minimum Off-Target Set Cover (MOTSC). Given a disease gene set, BTSC seeks a balanced solution that maximizes the coverage on the disease genes and minimizes the off-target hits at the same time. MOTSC seeks a full coverage on the disease gene set while minimizing the off-target set. Through simulation, both BTSC and MOTSC demonstrated a much faster running time over exhaustive search with the same accuracy. When applied to real disease gene sets, our algorithms not only identified known drug combinations, but also predicted novel drug combinations that are worth further testing. In addition, we developed a web-based tool to allow users to iteratively search for optimal drug combinations given a user-defined gene set.

Availability: Our tool is freely available for noncommercial use at http://www.drug.liuzlab.org/.

Contact: zhandong.liu@bcm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 19, 2013; revised on December 31, 2013; accepted on January 21, 2014

1 INTRODUCTION

Complex diseases, such as cancer, cardiovascular diseases and neurological disorders, usually involve multiple genes whose protein products play pivotal roles in controlling aberrant pathways and networks. In addition, disease pathways and networks are often redundant and robust to single-point perturbations. Because most drugs are designed to selectively target specific proteins, single-drug treatments usually cannot break down the whole disease pathways and networks. This is why the traditional 'one disease, one gene, one drug' treatment often fails (Hopkins, 2008). Multi-target treatments, especially drug combinations, can simultaneously target multiple components of the disease pathways and networks, thus offering hope for treating such complex diseases (Jia *et al*., 2009). There have already been many successful combinatorial therapies to treat complex diseases. For example, highly active antiretroviral therapy is a potent combination of at least three active antiretroviral drugs targeting reverse transcriptase, protease and integrase to keep the HIV virus from replicating itself (Lucas *et al*., 1999). The combination of glyburide and metformin is used to treat type 2 diabetes in complementary ways (Bokhari *et al*., 2003). Moduretic is the combination of amiloride and hydrochlorothiazide that can effectively treat patients with hypertension (HTN) (Frank, 2008; Wilson *et al*., 1988). The combination of anastrozole and fulvestrant is more effective than individual or sequential usage of both drugs for the treatment of hormone-receptor-positive metastatic breast cancer (Mehta *et al*., 2012).

Despite the increasing successes in using drug combinations to treat complex diseases, most of them were developed based on clinical experience or test-and-trial strategy, which is not only time-consuming but also expensive. High-throughput screening methods have also been developed to identify effective pairwise drug combinations (Borisy *et al*., 2003; Lehár *et al*., 2009; Tan *et al*., 2012). However, a systematic analysis of all the possible pairwise combinations is both labor-intensive and cost-ineffective because of the large combinatorial space needed to explore. Furthermore, most drug combinations may not significantly improve the efficacy over individual drugs. Therefore, large-scale drug combination screening is highly ineffective. In addition, a systematic screening becomes unfeasible if combinations of more than two drugs are considered. The closed-loop control (Wong *et al*., 2008), stack sequential (Calzolari *et al*., 2008) and other search algorithms reviewed in Feala *et al*. (2010) have been developed together with biological experiments to identify the optimal drug combinations from a huge drug-dose space. However, given thousands of individual drugs, careful preselection of a subset of drugs within which to search

*To whom correspondence should be addressed.

for optimal combinations in experimental testing is not an easy task.

Several systematic computational approaches for predicting drug combinations have recently been developed and could provide a guide to limit the search space for experimental methods. For example, Vazquez (2009) proposed identifying the optimal drug combinations by searching for the minimal number of drugs that can target all the cancer cell lines using the highest-degree-first and simulated annealing algorithms. However, this heuristic search algorithm does not have optimum guarantee, and the convergence rate for large-scale datasets could be slow. Wu *et al*. (2010) integrated a molecular interaction network and gene expression data of individual drugs to identify subnetworks affected by individual or combinatorial drugs. The drug effect on these subnetworks is measured by taking into account both efficacy and side effect and then used to prioritize drug combinations. They successfully identified effective drug combinations used to treat type 2 diabetes, but the dependency on the availability of gene expression data treated with individual and combinatorial drugs as well as the high computational cost when handling the vast combinatorial space makes their approach unsuitable for large-scale application. Zhao *et al*. (2011) and Xu *et al*. (2012) proposed two similar computational approaches to prioritize pairwise drug combinations using feature patterns enriched in the known drug combinations and got some promising predictions. However, their approaches rely heavily on the known drug combination data that are of small size, thus biasing their predictions toward those combinations that are similar to the known ones. Therefore, the current computational approaches are limited, and there is a great need to develop new algorithms for drug combination prediction.

Drug design is usually specific, but one drug may target multiple proteins due to promiscuous binding (Paolini *et al*., 2006; Yildirim *et al*., 2007). One protein can be targeted by multiple drugs as well. The drugs and proteins form an intricate drug–target network. When used to treat diseases, a drug can affect both disease on-target proteins for which it was designed and some off-target proteins that are not related to the diseases. When different drugs are combined, there could be a large number of additive off-targets whose effects are undesired. Thus, given a set of disease genes, the problem of finding the optimum drug combinations that maximize on-target coverage and minimize off-target effects is important and challenging. This is similar to the efficacy maximization and side effect minimization problem in combinatorial therapy. Also, such off-target effects in drug combination have not been considered in the systematic computational studies (Vazquez, 2009; Xu *et al*., 2012; Zhao *et al*., 2011), with the exception of the work (Wu *et al*., 2010).

To address the on-target maximization and off-target minimization problem, we first formulated the optimal combinatorial therapy problem using an optimization framework and solved it using mixed integer linear programming (MILP). Then, we compared our approach with exhaustive search using simulation and demonstrated a better performance of our approach over exhaustive search. Finally, we demonstrated the good performance of our approach through searching optimal drug combinations for six disease gene sets from a drug–target network. Our approach not only captured the well-known drugs and drug combinations, but also suggested novel uses for some other drugs. The drug combinations discovered using our approach can target the protein products of disease genes with minimal perturbations to the other proteins and are worthy of further testing to treat such diseases. To make our approach more accessible to the general drug discovery community, we developed a web-based tool to allow users to iteratively search for optimal drug combinations from a user-defined gene set.

The remaining part of this article is organized as follows: we first describe the optimization approach in Section 2; application of the approach on both the simulated and real data is presented in Section 3; and a brief discussion on current issues and potential future improvements is described in the concluding section of the article.

2 METHODS

2.1 Drug–target network

We extracted the drug–target interactions from the DrugBank database (version 3.0) (Knox *et al*., 2011) and constructed a bipartite network, in which nodes represent drugs or targets and edges represent drug–target interactions. We further removed the targets with no human gene symbol annotation. The remaining network contains 4233 drugs, 2058 target genes and 9669 drug–target interactions. We also extracted the adverse drug–drug interaction effects and the drug action information from the DrugBank database. Given an input disease gene set, drugs with the same set of targets, actions and interacting drugs were merged into a single meta-drug, as they are equivalent to our algorithm. Details on drug data processing are provided in Supplementary Text S1. Unless otherwise specified, we used the term drug instead of meta-drug in the remaining text and the term disease gene to represent the protein product of disease gene.

2.2 Optimal combinatorial therapy

Given a set of disease genes, we first removed those genes that have no associated drugs in the drug–target network. The remaining set of disease genes is called on-target set, denoted as $T = \{t_1, t_2, \ldots, t_p\}$. We then extracted the drugs associated with the on-target genes from the drug–target network. The set of off-targets, $S = \{s_1, s_2, \ldots, s_q\}$, is the set of genes that are connected with the on-target-associated drugs in the drug–target network, but does not overlap with T. Mathematically, the associated drugs can be formulated as $\mathcal{M} = \{M_i | M_i \subseteq T \cup S, i = 1, 2, \ldots, m\}$, where each drug M_i can target some disease genes in T and some off-target genes in S as well. Finding the optimal drug combination for a disease is to maximize the coverage on T and minimize the overlap with S using a subset of $\mathcal{M}$. This problem can be defined as follows.

Problem 1.

Balanced Target Set Cover (BTSC) problem. Given a disease $D = (T, S)$ and a collection of drugs $\mathcal{M}$, find a subset $\mathcal{C} \subseteq \mathcal{M}$ and $|\mathcal{C}| \leq k$ that minimize the $cost(D, \mathcal{C}) = \alpha|T \setminus (\cup\mathcal{C})| + (1 - \alpha)|S \cap (\cup\mathcal{C})|$, where $\cup\mathcal{C} = \cup_{C \in \mathcal{C}} C$, k is the upper bound on the cardinality of the solution set $\mathcal{C}$ and α is the weight balance of the coverage between on-target set T and off-target set S.

A natural choice of α is 0.5. When α is set to 1, BTSC is equivalent to the maximum coverage problem. In practice, users may have a small set of highly confident drug targets and would require full coverage on these targets while minimizing the number of off-targets. BTSC cannot be used directly to solve this problem.

To overcome this limitation, we further propose to fully cover the on-targets in T and minimize the number of off-targets in S. In other words,

we require any selected combination of drugs to cover all the on-target disease genes. This can be achieved by setting α to 0 and adding a constraint of full set cover. Mathematically, the restricted problem can be defined as follows:

PROBLEM 2.

Minimum Off-Target Set Cover (MOTSC) problem. Given a disease $D = (T, S)$ and a collection of drugs $\mathcal{M}$, find a subset $\mathcal{C} \subseteq \mathcal{M}$ that minimizes the $cost(D, \mathcal{C}) = |S \cap (\cup \mathcal{C})|$, where $|T \cap (\cup \mathcal{C})| = |T \cap (\cup \mathcal{M})|$ and $\cup \mathcal{C} = \cup_{C \in \mathcal{C}} C$.

2.3 BTSC and MOTSC problems are NP-hard

The NP-hard property of the BTSC problem can be proved through reduction mapping. By setting α to 1, the maximum coverage problem becomes finding a maximum cover on the disease genes in T using at most k drugs from $\mathcal{M}$. This reduction shows that BTSC is at least as hard as the maximum coverage problem, which is in the NP-hard problem set (Cohen and Katzir, 2008).

MOTSC can be viewed as a modified version of the Red-Blue Set Cover, which is also a generalization of the standard set cover problem (Miettinen, 2009). Red-Blue Set Cover problem is much harder than the standard set cover problem, and there exists no polynomial approximation with a factor of $2^{(4 \log n)^{1-\epsilon}}$ for any $\epsilon > 0$ (Peleg, 2007).

2.4 Mathematical programming formulation

Mathematically, BTSC can be formulated using MILP as follows:

$$\text{minimize } \alpha \sum_{i=1}^{p} (1 - y_i) + (1 - \alpha) \sum_{i=p+1}^{p+q} y_i \tag{1}$$

$$\text{subject to } (\mathbf{Bx})_i - y'_i = 0 \tag{2}$$

$$by_i - y'_i \geq 0 \tag{3}$$

$$y_i - y'_i \leq 0 \tag{4}$$

$$\sum_{j=1}^{m} x_j \leq k \tag{5}$$

$$(\mathbf{Lx}) \preceq 1 \tag{6}$$

$$y'_i \in \mathbb{Z}^+, y_i, x_j \in \{0, 1\} \tag{7}$$

The formulation has three variables, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{y}'$. The first two are binary, and the last one is nonnegative. The binary solution vector $\mathbf{x}$ indicates which drugs are selected. The nonnegative cost variable y'_i counts the times that the i_{th} gene is covered by the selected drugs. The binary cost vector $\mathbf{y}$ is derived from $\mathbf{y}'$, $y_i = 1$ if $y'_i \geq 1$; otherwise, $y_i = 0$. The value of y_i indicates whether the i_{th} gene is covered. We note that in vectors $\mathbf{y}$ and $\mathbf{y}'$, the first p elements are associated with the on-target genes and the next q elements are associated with the off-target genes.

The relation between a given disease D and the associated drugs $\mathcal{M}$ can be represented using a binary matrix $\mathbf{B}$, where the rows are indexed by the on-targets (p) and off-targets (q), and the columns represent the drugs. $\mathbf{B}_{im} = 1$ if the i_{th} gene in D is covered by the m_{th} drug in $\mathcal{M}$; otherwise, $\mathbf{B}_{im} = 0$. The nonzero elements in the product vector of $\mathbf{B}$ and $\mathbf{x}$ indicate the corresponding genes targeted by a selection of drugs.

The drug–drug adverse interaction effects are encoded in $\mathbf{L}$, a $l \times m$ binary matrix where, for each row k, $\mathbf{L}_{ki} = \mathbf{L}_{kj} = 1$ if the drugs M_i and M_j have an adverse effect when used together.

The intuition of the MILP formulation is as follows. The equality constraint (2) counts the number of times that an on-target or an off-target is covered by the selected drug combination. However, $\mathbf{y}'$ cannot be directly used in the objective cost function (1). Thus, the binary cost $\mathbf{y}$ is introduced and related to $\mathbf{y}'$ by the inequality constraints (3) and (4). The inequality constraint (3) requires $y_i = 1$ if $y'_i \geq 1$. For that, the value of b needs to be at least the maximum value of all y'_i. The inequality constraint (4) guarantees that $y_i = 0$ whenever $y'_i = 0$. The maximum number of drugs for any feasible solution is bounded by the inequality constraint (5). The inequality constraint (6) guarantees any feasible solution to avoid the drug–drug adverse effects encoded in $\mathbf{L}$. In this article, we solved BTSC problem using the GNU Linear Programming Kit package (http://www.gnu.org/software/glpk/). MOTSC can be formulated similarly (Supplementary Text S2). The dual problems of BTSC and MOTSC are further discussed in the Supplementary Texts S3 and S4.

2.5 Iterative search and online tool

To make our approach more accessible in practice, we developed a web-based interactive tool (http://www.drug.liuzlab.org/) that will allow users to iteratively refine the search results. Both BTSC and MOTSC are implemented with the options to filter out drugs based on approval status, drug–drug adverse interaction and drug action direction. In addition, our web tool can also generate the output result files for Cytoscape visualization (Shannon *et al.*, 2003).

3 RESULTS

3.1 Time complexity and accuracy analysis

To evaluate the time complexity and accuracy of our algorithm, we compared it with exhaustive search (ES) on a simulated dataset. According to the formulation in Section 2.4, the simulated dataset is controlled by four parameters: (i) the number of columns in $\mathbf{B}$, which represents the number of drugs associated with a disease gene set; (ii) the number of rows in $\mathbf{B}$, which represents the total on-target and off-target space; (iii) the density of $\mathbf{B}$, which represents the percentage of genes targeted by a drug; and (iv) the value of p, which represents the number of on-targets.

The first parameter, the number of columns in $\mathbf{B}$, is the most important factor affecting the running time of ES. Therefore, we generated $\mathbf{B}$ by varying the column size from 10 to 30 with a step size equal to 2. At the same time, the second parameter was set to 1000, and $\mathbf{B}$ was sampled from a Bernoulli distribution with hitting probability equal to 0.01 (the third parameter). In addition, we further removed those rows in $\mathbf{B}$ for which the sum across the columns is equal to 0. We sampled p indexes (the third parameter) of the rows in $\mathbf{B}$, with a sampling rate equal to 10%. Then, we simulated the data 10 times with each varying number of the columns in $\mathbf{B}$, applied BTSC and ES to find the solution $\mathbf{x}$ and compared their differences in the cost and running time.

Our results demonstrated that there is no difference in the cost function between the BTSC and ES solutions. However, the running time of ES increases exponentially, while BTSC tends to generate the correct results significantly faster (Fig. 1). From these results, we found that ES is impractical even for finding combinations among a small number of drugs. For example, a search for combinations among 30 drugs using ES will take ~4.7 days to get the optimal solution. However, BTSC can obtain the optimal solution within 9 s. We also varied the other three parameters in simulation. There is still no cost difference between BTSC and ES, and the effects of these three parameters on the running time difference are much less important than that of the

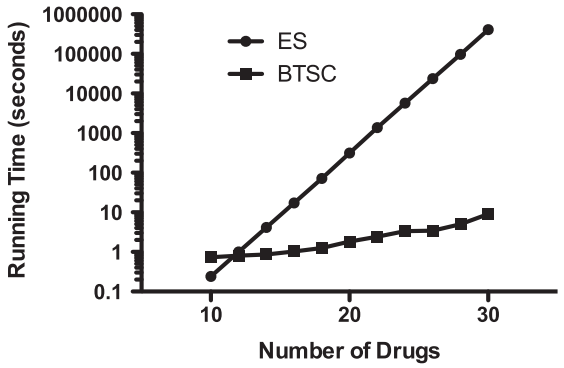

Fig. 1. Running time difference between BTSC and ES. The number of associated drugs is simulated from 10 to 30 with a step size equal to 2. Each data point represents the average running time of 10 replicates. For the simulated data with 30 associated drugs, ES takes ~4.7 days, whereas BTSC needs only 9 s

first parameter (data not shown). Similar results were obtained on MOTSC (Supplementary Fig. S1). Thus, through this simulation, we demonstrated that our algorithms can achieve high accuracy with much faster running time.

To demonstrate the power of our approach in searching for optimal drug combinations, we applied it on six disease gene sets: acute myocardial infarction (AMI), EGFR-PI3K-AKT-mTOR (EPAM) pathway, HTN, type 2 diabetes mellitus (T2DM), Parkinson's disease (PD) and schizophrenia (SZ). All the search parameters used for the following results are available in the Supplementary Text S5. All the following networks are visualized using Cytoscape (Shannon *et al.*, 2003). Drugs are represented using DrugBank names, except that drugs with long names are represented using DrugBank IDs. Further details are provided in the online document.

3.2 Acute myocardial infarction

AMI is the irreversible necrosis of heart muscle with the usual cause that a coronary artery is obstructed by an acute thrombus. The AMI-related 20 genes were derived from BioCarta AMI pathway (http://www.biocarta.com), which is collected in the Molecular Signatures Database (version 3.1) (Liberzon *et al.*, 2011; Subramanian *et al.*, 2005). From 35 associated drugs, BTSC predicted that streptokinase, fondaparinux sodium|enoxaparin, DB04134 and DB07376 can be combined to target the AMI gene set (Fig. 2).

In the four-drug combination, streptokinase is able to dissolve blood clots that have formed in the blood vessels. Fondaparinux and enoxaparin are anticoagulants that help prevent the formation of blood clots. The coadministration of streptokinase and fondaparinux in a published cohort of ST-elevation myocardial infarction patients significantly reduced the risk of death, recurrence and severe bleeding (Peters *et al.*, 2008). The combination of streptokinase and enoxaparin is effective in patients with AMI (Giraldez *et al.*, 2009; Simoons *et al.*, 2002).

Two experimental compounds, DB04134 and DB07376, were also identified in the solution. DB04134 is an antifibrinolytic agent that acts by inhibiting plasminogen activators that have fibrinolytic properties. DB07376 is an anticoagulant that can

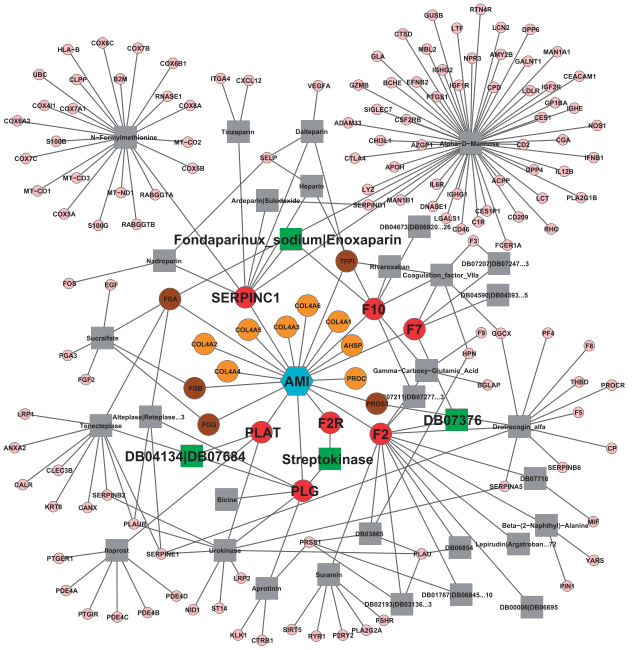

Fig. 2. The disease–gene–drug network of AMI. The disease gene set name AMI is represented using a cyan hexagon, and there are 20 genes (red, orange or chocolate circles) in the AMI gene set. Four (green squares) of 35 associated drugs (green or gray squares) selected by BTSC can cover seven AMI genes (red circles) with no known off-target. Eight genes (orange circles) have no associated drugs, and five genes (chocolate circles) associated with drugs are not covered by the selected drug combination

prevent blood clotting. Because these two compounds target different AMI genes, the use of these two compounds in combination with streptokinase and/or fondaparinux sodium|enoxaparin may exert additional benefits in treating AMI patients.

This result indicated that BTSC is able to identify known drug combinations as well as novel compounds that are not covered by existing therapies.

3.3 EPAM pathway

EPAM pathway plays an important role in a lot of cellular processes and is abnormally activated in a variety of human diseases including cancer and neurological disorders. Eighteen genes in the EPAM pathway were obtained from Hennessy *et al.* (2005) and Morris *et al.* (2011). From 41 associated drugs, BTSC predicted that lapatinib, everolimus, DB08059, DB07812 and DB06831 can be combined to target the EPAM pathway (Fig. 3).

In the five-drug combination, lapatinib is a selective inhibitor developed for both EGFR and ERBB2 (Diaz *et al.*, 2010; Opdam *et al.*, 2012). Everolimus, a derivative of rapamycin, can act as an mTOR inhibitor and block the signal transduction for cell proliferation. Through literature search, we discovered that the combination of lapatinib and everolimus shows efficacy in patients with advanced cancers in a phase I study (Gadgeel *et al.*, 2013). This combination is also under a clinical

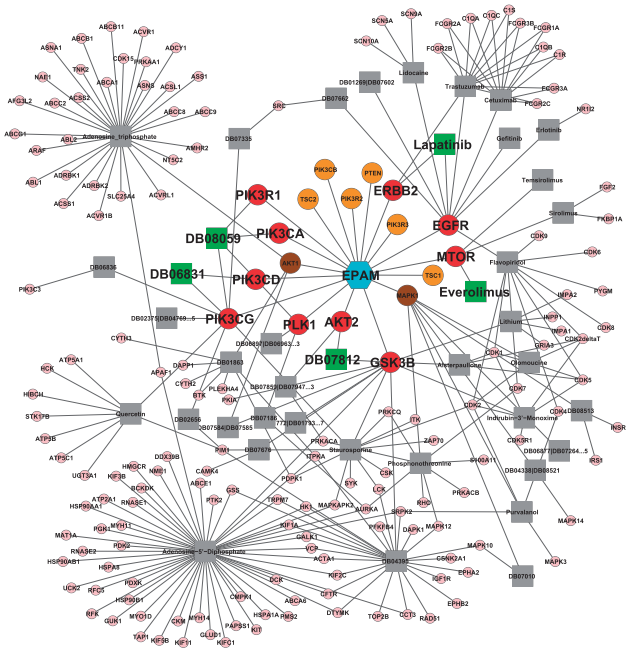

Fig. 3. The disease–gene–drug network of EPAM. The disease gene set name EPAM is represented using a cyan hexagon, and there are 18 genes (red, orange or chocolate circles) in the EPAM gene set. Five (green squares) of 41 associated drugs (green or gray squares) selected by BTSC can cover 10 EPAM genes (red circles) with no known off-target. Six genes (orange circles) have no associated drugs, and two genes (chocolate circles) associated with drugs are not covered by the selected drug combination

trial (NCT01272141) in patients with advanced triple-negative breast cancer and under a phase 1b/2 clinical trial (NCT01783756) for treatment of HER-2 positive breast cancer with central nervous system metastasis.

One consequence of mTOR inhibition by rapamycin or its derivatives is the loss of inhibition on PI3K and AKT, resulting increased pathway activity. In addition, patients treated with EGFR inhibitors often develop drug resistance through secondary mutations (Kobayashi *et al*., 2005). To address these issues, our combinatorial therapy selected DB08059 (wortmannin), a known inhibitor of PI3K, and an experimental inhibitor (DB07812) of AKT and GSK3β. Therefore, simultaneously targeting multiple genes in the EPAM pathway will offer the hope to increase the efficacy of drug treatment and delay the development of drug resistance.

This result indicated that BTSC is effective in identifying optimal drug combination to target multiple components of an aberrant pathway.

3.4 Hypertension

HTN or high blood pressure is a chronic medical condition characterized by the elevated blood pressure in the arteries. Yue *et al*. (2006) curated a set of 26 HTN-related genes including key regulators in the renin–angiotensin pathway, endothelin regulation, natriuretic peptide regulation and bradykinin–kallikrein pathway. From 77 associated drugs, BTSC predicted that the combination of bupranolol, triamterene, chlorthalidone, sitaxentan and remikiren could be effective for treating HTN (Fig. 4).

In the five-drug combination, bupranolol is a nonselective beta blocker, triamterene is a potassium-sparing diuretic, and chlorthalidone is a thiazide diuretic. The combination of bupranolol and triamterene has been well documented as an effective agent for HTN treatment (Schrey, 1981). The combination of triamterene and chlorthalidone shows clinical efficacy in reducing blood pressure, and keeping serum potassium concentration from decreasing too much (Spiers and Wade, 1996).

Two other HTN drugs, sitaxentan and remikiren, were also identified in the solution. Sitaxentan is an endothelin receptor antagonist, and remikiren is a renin inhibitor. Because these two drugs target different pathways, they may provide additional benefits for patients who are not responsive to the above drug cocktails.

This result indicated that BTSC is also effective in identifying optimal drug combination to target multiple altered pathways.

3.5 Type 2 diabetes mellitus

T2DM is a chronic metabolic disorder in which blood glucose is increased to a high level due to altered insulin signaling. We extracted Kyoto Encyclopedia of Genes and Genomes insulin signaling pathway (http://www.genome.jp/kegg/) and REACTOME insulin receptor signaling cascade pathway (http://www.reactome.org) from the Molecular Signatures Database (version 3.1) (Liberzon *et al*., 2011; Subramanian *et al*., 2005). These two pathways share 41 genes, which were used as T2DM-related genes for optimal drug combination search. From 13 associated drugs, BTSC predicted that insulin aspart|insulin detemir, metformin, everolimus|temsirolimus and pegademase bovine could be combined to treat T2DM (Supplementary Fig. S2).

In the four-drug combination, insulin aspart is a fast-acting insulin analog, while insulin detemir is a long-acting insulin analog. Metformin can help control blood sugar levels. Insulin including insulin aspart and insulin detemir in combination with metformin is routinely used in the treatment of T2DM (Hollander *et al*., 2011; Kvapil *et al*., 2006; Wulffelé *et al*., 2002).

Inhibition of mTOR leads to upregulation of insulin receptor substrate and increased activity of AKT (O'Reilly *et al*., 2006), thus being able to ameliorate insulin resistance. Everolimus and temsirolimus, two inhibitors of mTOR selected by BTSC, are worthy of further testing in combination with insulin and/or metformin to combat T2DM.

This result demonstrated that BTSC has the ability to identify well-known combination regimen regularly used in disease management and promising combination component with additional beneficial effect.

3.6 Parkinson's disease

PD is a movement disorder caused by depletion of brain dopamine. We obtained 10 genes involved in dopamine synthesis and metabolism pathway from Youdim *et al*. (2006). From 28 associated drugs, BTSC predicted that the combination of levodopa, carbidopa, entacapone, selegiline and metyrosine is able to fully cover PD-related genes (Supplementary Fig. S3).

Among the five selected drugs, there are several well-known combinations that have already been used to treat PD. For

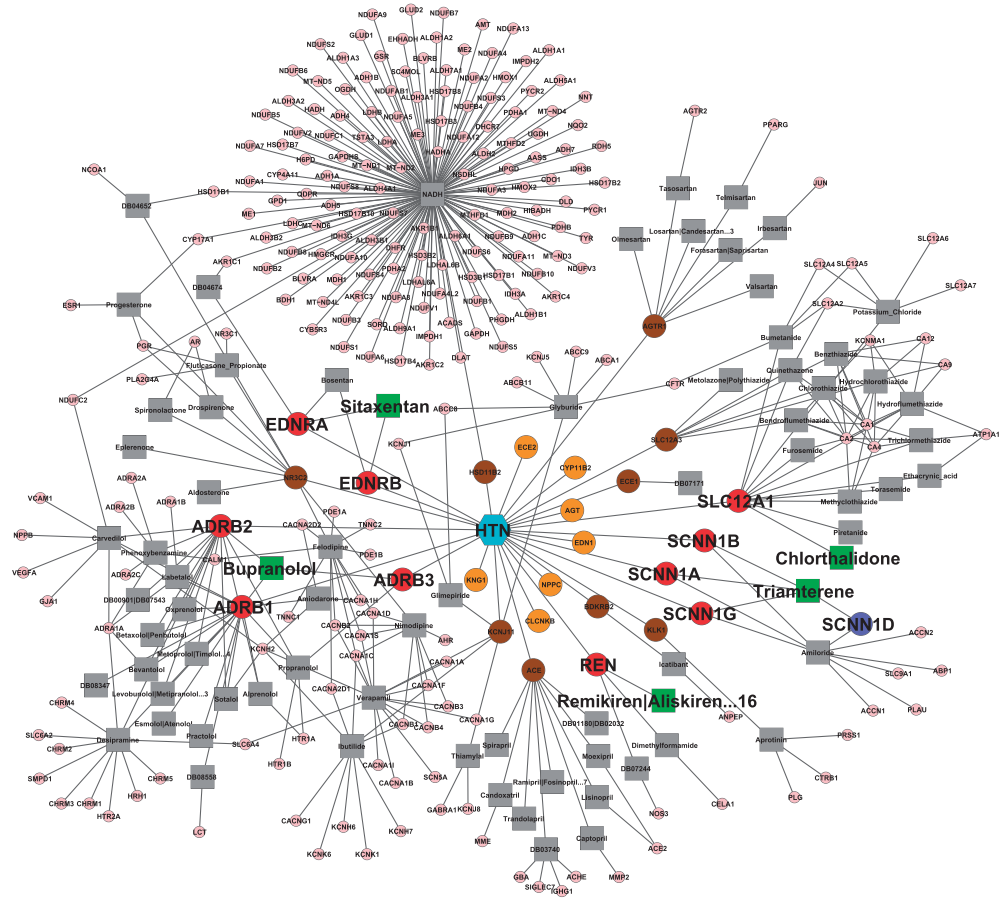

Fig. 4. The disease–gene–drug network of HTN. The disease gene set name HTN is represented using a cyan hexagon, and there are 26 genes (red, orange or chocolate circles) in the HTN gene set. Five (green squares) of 77 associated drugs (green or gray squares) selected by BTSC can cover 10 HTN genes (red circles) with only one off-target (blue circle). Seven genes (orange circles) have no associated drugs, and nine genes (chocolate circles) associated with drugs are not covered by the selected drug combination

example, the triple combination of levodopa (a dopamine replacement), carbidopa (a dopa decarboxylase inhibitor) and entacapone (a catechol-O-methyltransferase inhibitor) has been approved by Food and Drug Administration (FDA) for treating PD (Hauser, 2004). In addition, selegiline (a monoamine oxidase inhibitor) in combination with levodopa, carbidopa or entacapone shows improved treatment of PD (Cedarbaum *et al*., 1991; Elizan *et al*., 1991; Lyytinen *et al*., 1997).

This result further demonstrated the good performance of BTSC in identifying optimal drug combination.

3.7 Schizophrenia

SZ is a chronic, severe and disabling brain disorder affecting ~1% of the population worldwide and is linked to the dysfunction of dopamine, gamma-aminobutyric acid, glutamate and serotonin pathways. We extracted 28 SZ-related receptor and transport genes involved in these four pathways from the Schizophrenia Gene Resource database (Jia *et al*., 2010; Sun *et al*., 2008, 2009).

From 102 associated drugs, BTSC predicted that haloperidol, sertraline, dolasetron, methohexital and vilazodone could be combined to treat patients with SZ (Supplementary Fig. S4). To validate the prediction, we did a literature search and discovered that the coadministration of haloperidol and sertraline to schizophrenics results in clinical improvement of negative symptoms and aggravation of extrapyramidal symptoms (Lee *et al*., 1998).

BTSC also unveiled possible new uses of three FDA-approved drugs that have not been previously used in SZ. In particular, dolasetron targeting the serotonin pathway has been used to treat nausea and vomiting following chemotherapy. Because the serotonin pathway plays a key role in mood control, this drug may improve the existing treatment schema. Such drugs could be reused to target SZ in a cost-effective way because their safety and side effects have already been well studied.

This result not only further demonstrated the good performance of BTSC in optimal drug combination prediction, but also suggested the potential utility of BTSC in drug repositioning.

3.8 Performance on DCDB

We further performed an independent validation of BTSC by applying it to predict the FDA-approved drug combinations in

the Drug Combination Database (DCDB) (Liu *et al*., 2010). Using the FDA-approved drug combinations derived from DCDB, we first generated an approximately gold standard dataset of 68 known associations between disease gene sets and drug combinations. We then performed search to see the extent to which these approved drug combinations can be recovered by our online tool. Totally, 59 of 68 (recovery rate = 86.8%) approved drug combinations can be fully or partly recovered, whereas 55 of 68 (recovery rate = 80.9%) approved drug combinations can be fully recovered. Details are provided in the Supplementary Text S6. This result demonstrated that BTSC is able to identify known drug combinations with high accuracy in a large scale.

Taken together, these results indicated that BTSC has good performance in predicting both known and novel drug combinations. In addition, MOTSC identified similar, but slightly different, drug combinations when applied to the five disease gene sets that are not fully covered by BTSC (Supplementary Figs S5–S9). The difference between the results of BTSC and MOTSC further indicated that MOTSC is more applicable if the user has a specific requirement on the full coverage of the input genes. Thus, BTSC and MOTSC provided two complementary ways to identify optimal drug combination.

4 CONCLUSION

Optimal combinatorial therapy discovery is an important and challenging problem. In this article, we introduced two complementary approaches for this problem and solved them using MILP. There are many heuristic search algorithms that can provide greedy solutions for this problem. However, we are interested in finding the global optimum solution, and these heuristic search algorithms cannot provide such accuracy. Therefore, we excluded this category of algorithms in solving the optimal combinatorial therapy problem and did not compare our approaches with any heuristic search algorithms. Instead, we compared our approaches with exhaustive search and demonstrated that our algorithms can obtain the same optimum solution with much faster running time. Application of our approach on real disease gene sets demonstrated its good performance in identifying known drug combinations as well as predicting novel drug combinations. In addition, our approach has the potential to unveil new functions of existing drugs for drug repositioning.

We have developed an online tool for our proposed algorithms. The online tool provides many features, such as exclusion of adverse drug–drug interaction, constraint on the number of drugs, iterative search, highly efficient solver and email notification. In addition, the online tool allows the user to assign the weight balance on the on-target and off-target sets as well as to choose how to handle drugs with opposite actions on input genes. The online tool is also flexible to include new constraints, such as the importance weight of input disease genes and the penalty weight of potential off-target genes. The availability of the online tool will make our algorithm accessible not only to the computational biologists but also to the bench scientists.

The mathematical analysis presented here provides a general framework for the solution of multi-target therapeutic design. Although running our algorithm on arbitrarily large set of drugs is not possible due to the nature of NP-hard problem, most of the real applications fall into the small and mid-size categories. In practice, our algorithm and online tool demonstrate accurate and fast performance on those applications. One weakness of our approach stems from its use of incomplete drug–target interaction data. However, in recent years, a significant amount of effort has been devoted to drug target annotation and prediction. With the improved size and quality of such data, we believe that our approach and web-based tool will play an increasing role in drug discovery and development. In addition, personal variants in protein-coding genes can be easily obtained with the advances in next-generation sequencing technologies. Owing to the heterogeneity of complex diseases, even individual patients with the same disease may have a distinct set of causal genes and thus will need different treatment strategy. This problem is challenging and cannot be resolved in an effective way now. However, our tool offers one way to help predict optimal drug combination for targeting individual set of disease genes, which will have a non-trivial contribution to personalized medicine.

ACKNOWLEDGEMENTS

The authors extend their gratitude to Drs Juan Botas, Huda Zoghbi and James Alverez. The authors also thank the reviewers for their valuable suggestions on improving the manuscript.

Funding: This project is supported by Houston Bioinformatics Endowment and [National Institutes of Health (5DP5OD009134)]. Z.L. is also supported by [National Science Foundation/Division of Mathematical Sciences (1263932 in part)]. J.S. is supported by Cancer Prevention & Research Institute of Texas (R1307) Rising Star Award to Dr Hua Xu.

Conflict of Interest: none declared.

REFERENCES

Bokhari,S.U. *et al.* (2003) Beneficial effects of a glyburide/metformin combination preparation in type 2 diabetes mellitus. *Am. J. Med. Sci.*, **325**, 66–69.

Borisy,A.A. *et al.* (2003) Systematic discovery of multicomponent therapeutics. *Proc. Natl Acad. Sci. USA*, **100**, 7977–7982.

Calzolari,D. *et al.* (2008) Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput. Biol.*, **4**, e1000249.

Cedarbaum,J.M. *et al.* (1991) L-deprenyl (selegiline) added to Sinemet CR in the management of Parkinson's disease patients with motor response fluctuations. *Clin. Neuropharmacol.*, **14**, 228–234.

Cohen,R. and Katzir,L. (2008) The generalized maximum coverage problem. *Inf. Process. Lett.*, **108**, 15–22.

Diaz,R. *et al.* (2010) Antitumor and antiangiogenic effect of the dual EGFR and HER-2 tyrosine kinase inhibitor lapatinib in a lung cancer model. *BMC Cancer*, **10**, 188.

Elizan,T.S. *et al.* (1991) Early combination of selegiline and low-dose levodopa as initial symptomatic therapy in Parkinson's disease. Experience in 26 patients receiving combined therapy for 26 months. *Arch. Neurol.*, **48**, 31–34.

Feala,J.D. *et al.* (2010) Systems approaches and algorithms for discovery of combinatorial therapies. Wiley interdisciplinary reviews. *Syst. Biol. Med.*, **2**, 181–193.

Frank,J. (2008) Managing hypertension using combination therapy. *Am. Fam. Physician*, **77**, 1279–1286.

Gadgeel,S.M. *et al.* (2013) Phase I study evaluating the combination of lapatinib (a Her2/Neu and EGFR inhibitor) and everolimus (an mTOR inhibitor) in patients with advanced cancers: South West Oncology Group (SWOG) Study S0528. *Cancer Chemother. Pharmacol.*, **72**, 1089–1096.

Giraldez,R.R. *et al.* (2009) Streptokinase and enoxaparin as an alternative to fibrin-specific lytic-based regimens: an ExTRACT-TIMI 25 analysis. *Drugs*, **69**, 1433–1443.

Hauser,R.A. (2004) Levodopa/carbidopa/entacapone (Stalevo). *Neurology*, **62** (**1 Suppl. 1**), S64–S71.

Hennessy,B.T. *et al.* (2005) Exploiting the PI3K/AKT pathway for cancer drug discovery. *Nat. Rev. Drug Discov.*, **4**, 988–1004.

Hollander,P. *et al.* (2011) Efficacy and safety of insulin detemir once daily in combination with sitagliptin and metformin: the TRANSITION randomized controlled trial. *Diabetes Obes. Metab.*, **13**, 268–275.

Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.

Jia,J. *et al.* (2009) Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.*, **8**, 111–128.

Jia,P. *et al.* (2010) SZGR: a comprehensive schizophrenia gene resource. *Mol. Psychiatry*, **15**, 453–462.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Kobayashi,S. *et al.* (2005) EGFR mutation and resistance of non–small-cell lung cancer to Gefitinib. *N. Engl. J. Med.*, **352**, 786–792.

Kvapil,M. *et al.* (2006) Biphasic insulin aspart 30 plus metformin: an effective combination in type 2 diabetes. *Diabetes Obes. Metab.*, **8**, 39–48.

Lee,M.S. *et al.* (1998) Co-administration of sertraline and haloperidol. *Psychiatry Clin. Neurosci.*, **52** (**Suppl.**), S193–S198.

Lehár,J. *et al.* (2009) Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.*, **27**, 659–666.

Liberzon,A. *et al.* (2011) Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**, 1739–1740.

Liu,Y. *et al.* (2010) DCDB: drug combination database. *Bioinformatics*, **26**, 587–588.

Lucas,G.M. *et al.* (1999) Highly active antiretroviral therapy in a large urban clinic: risk factors for virologic failure and adverse drug reactions. *Ann. Intern. Med.*, **131**, 81–87.

Lyytinen,J. *et al.* (1997) Simultaneous MAO-B and COMT inhibition in L-Dopa-treated patients with Parkinson's disease. *Mov. Disord.*, **12**, 497–505.

Mehta,R.S. *et al.* (2012) Combination anastrozole and fulvestrant in metastatic breast cancer. *N. Engl. J. Med.*, **367**, 435–444.

Miettinen,P. (2009) *Matrix Decomposition Methods for Data Mining: Computational Complexity and Algorithms*. PhD thesis, University of Helsinki, Finland.

Morris,L.G.T. *et al.* (2011) Genomic dissection of the epidermal growth factor receptor (EGFR)/PI3K pathway reveals frequent deletion of the EGFR phosphatase PTPRS in head and neck cancers. *Proc. Natl Acad. Sci. USA*, **108**, 19024–19029.

Opdam,F.L. *et al.* (2012) Lapatinib for advanced or metastatic breast cancer. *Oncologist*, **17**, 536–542.

O'Reilly,K.E. *et al.* (2006) mTOR inhibition induces upstream receptor tyrosine kinase signaling and activates Akt. *Cancer Res.*, **66**, 1500–1508.

Paolini,G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.

Peleg,D. (2007) Approximation algorithms for the label-covermax and red-blue set cover problems. *J. Discrete Algorithms*, **5**, 55–64.

Peters,R.J.G. *et al.* (2008) OASIS-6 Investigators. The role of fondaparinux as an adjunct to thrombolytic therapy in acute myocardial infarction: a subgroup analysis of the OASIS-6 trial. *Eur. Heart J.*, **29**, 324–331.

Schrey,A. (1981) Hypertension treatment with beta blockers and diuretics. Treatment with a combination of bemetizid, triamterene and bupranolol. *Med. Welt*, **32**, 985–987.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Simoons,M. *et al.* (2002) AMI-SK Investigator. Improved reperfusion and clinical outcome with enoxaparin as an adjunct to streptokinase thrombolysis in acute myocardial infarction. The AMI-SK study. *Eur. Heart J.*, **23**, 1282–1290.

Spiers,D.R. and Wade,R.C. (1996) Double-blind parallel study of a combination of chlorthalidone 50 mg and triamterene 50 mg in patients with mild and moderate hypertension. *Curr. Med. Res. Opin.*, **13**, 409–415.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Sun,J. *et al.* (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics*, **25**, 2595–6602.

Sun,J. *et al.* (2008) Candidate genes for schizophrenia: a survey of association studies and gene ranking. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **147B**, 1173–1181.

Tan,X. *et al.* (2012) Systematic identification of synergistic drug pairs targeting HIV. *Nat. Biotechnol.*, **30**, 1125–1130.

Vazquez,A. (2009) Optimal drug combinations and minimal hitting sets. *BMC Syst. Biol.*, **3**, 81.

Wilson,D.R. *et al.* (1988) Interaction of amiloride and hydrochlorothiazide with atrial natriuretic factor in the medullary collecting duct. *Can. J. Physiol. Pharmacol.*, **66**, 648–654.

Wong,P.K. *et al.* (2008) Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm. *Proc. Natl Acad. Sci. USA*, **105**, 5105–5110.

Wu,Z. *et al.* (2010) A systems biology approach to identify effective cocktail drugs. *BMC Syst. Biol.*, **4** (**Suppl. 2**), S7.

Wulffelé,M.G. *et al.* (2002) Combination of insulin and metformin in the treatment of type 2 diabetes. *Diabetes Care*, **25**, 2133–2140.

Xu,K.J. *et al.* (2012) The drug cocktail network. *BMC Syst. Biol.*, **6** (**Suppl. 1**), S5.

Yildirim,M.A. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Youdim,M.B.H. *et al.* (2006) The therapeutic potential of monoamine oxidase inhibitors. *Nat. Rev. Neurosci.*, **7**, 295–309.

Yue,P. *et al.* (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.

Zhao,X.M. *et al.* (2011) Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Comput. Biol.*, **7**, e1002323.