Predicting Effective Drug Combinations via Network Propagation

Balázs Ligeti*, Roberto Vera†, Gergely Lukács*, Balázs Győrffy‡ and Sándor Pongor*†
*Faculty of Information Technology and Bionics, Pázmány Péter Catholic University,
Práter utca 50/A, H-1083 Budapest, Hungary; email: ligeti.balazs@itk.ppke.hu, lukacs@itk.ppke.hu
† Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology,
Padriciano 99, I-34012 Trieste, Italy; email: pongor@icgeb.org
‡ Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences,
Bókay u. 53-54, H-1083 Budapest, Hungary

***Abstract*— Drug combinations are frequently used in treating complex diseases including cancer, diabetes, arthritis and hypertension. Most drug combinations were found in empirical ways so there is a need of efficient computational methods. Here we present a novel method based on network analysis which estimates the efficacy of drug combinations from a perturbation analysis performed on a protein-protein association network. The results suggest that those drugs are likely to form effective combinations that perturb a large number of proteins in common, even if the original targets are found in seemingly unrelated pathways.**

I. Introduction

In the past few decades the number of new registered drugs has fallen much below the expectations despite novel technologies and growing investment in this area [1], [2]. Drugs designed by one drug - one target drug design strategy often fail at phase II or phase III of clinical trials not only because their side effects but also because of their insufficient therapeutic effects [1]. The latter problem is often attributed to the well-known robustness of biological systems against various kind of perturbations such as toxins, chemical compounds, mutations [2]. For instance, biological pathways are often redundant, diverse and modular as well as are rich in regulatory loops that can compensate the effect of perturbations. Multitarget drugs or drug combinations offer intriguing possibilities as they attack biological systems at multiple points which decreases the chances of compensatory effects. Simply put, a multiple attack makes the regrouping of resources more problematic. Ágoston et al. showed that multiple but partial knockout of targets is more efficient than a single but complete knockout [3]. In addition, drug combinations are known to have lower toxicity and higher therapeutic selectivity [4]. One is tempted to argue that multitarget drugs and drug combinations are a promising paradigm for drug development, the question is how the growing body of various databases can be utilized for this purpose [1], [5]. Even though the number of approved drug combinations is increasing, most of these were found by experience and intuition rather than in silico predictions [6]. Several experimental methods, even high throughput methods have been developed for measuring the efficiency of drug combinations [7]. This kind of exhaustive search may prove impractical, Wong et al. used a stochastic search algorithm to find the best combinations [8]. Yang et al. used differential equations to find a perturbation pattern that can return the system to a normal state from a disease state [9]. A common feature of these methods is that they require a very large number of experiments as well as a deep knowledge of the kinetic parameters of the pathways. Other methods use data mining algorithms to integrate pharmacological and network data [10], [11]. In this paper we present a novel drug combination prediction algorithm which is based on the assumption that the perturbations generated by the drugs propagate through the possible interactions between proteins and that drugs perturbing a large number of proteins in common can form effective combinations.

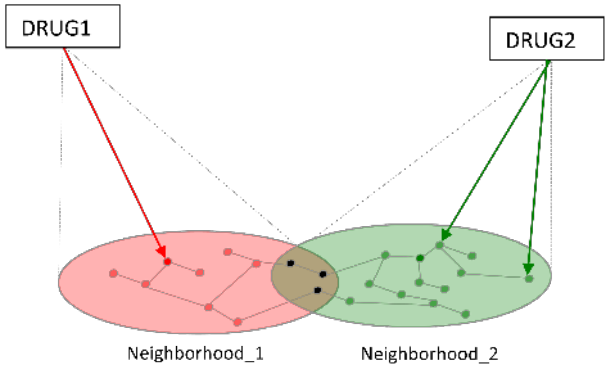

Fig. 1. *The network-interaction hypothesis. The effect of two drugs (Drug1, Drug2) first reaches their imminent targets (arrows) and the effect will propagate to network neighborhoods (subnetworks) indicated in red and green, respectively. The targets in the overlap are affected by both drugs, and we suppose that drugs affecting a large number of common targets will increase the effects of each other [12].*

II. Method and Experimental Design

A drug molecule affects its targets by various mechanisms including inhibition and activation. As the proteins are linked by complex networks of interactions, we can suppose that perturbation on a single protein will also propagate within the network. The propagation of an effect within a network can be described by random walk or diffusion models (**Figure 1**). The underlying – and admittedly speculative – hypothesis

is that one compound may not sufficiently reach the critical targets, but two pharmacons have greater chance. Our basic assumption is that the existing protein interaction network can provide a basis to find the proteins or genes perturbed by drug combination. In other words, we consider a drug combination effective if the network neighborhoods of the constituent pharmacons overlap (**Figure 1**). In order to find the commonly affected proteins, we use the PageRank algorithm [13], which has been successfully used to prioritize disease candidate genes [5], [14]. In this paper we only consider binary combinations. The interaction network is a graph $G(V, E)$ where V, E are the set of nodes and edges, respectively. In our case the nodes represent genes or proteins, and the edges are the associations or interactions between them. The edges may have a weight, which can be interpreted as an association strength. Let A be the adjacency matrix of the graph. The element a_{ij} is the weight of the edge between node i and j, if there is no edge then it is 0. One could define a random walk on that graph by rescaling the edges to transition probabilities. Let M be a stochastic matrix of the graph $G(V, E)$, then m_{ij} is the probability of going to node j from node i.

$$M = D^{-1}A$$

where D is a diagonal matrix:

$$D = diag(d_1, d_2, \ldots, d_N) \tag{1}$$

where $d_i = \sum_{j=1}^{|V|} A_{ij}$.

$$P^{k+1} = M^T P^k = (M^T)^k P^0$$

where P^k is a probability distribution, so p_i^k is the probability of being at node i in the step k. P^0 is the initial probability distribution vector, which are the probabilities of starting the random walk at a given a node.

a) PageRank: PageRank with prior [13] is a modified random walk, where in each step the random walker jumps back to one of the initial nodes or continues the traveling with a certain probability.

$$P^{(i+1)} = (1-\alpha)\left(M^T P^{(i)}\right) + \alpha P^0 \tag{2}$$

$$p_i^0 = \begin{cases} \frac{1}{|N_T|}, & \text{if the protein } i \text{ is drug target} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where N_T is the number of drug targets.

A. Randomizations and the network neighborhood of drugs

In protein interaction network there are nodes which are more central, i.e. have a higher degree, and are more likely to be reached by chance. In order to avoid this bias, randomization procedure was applied to estimate the statistical significance of each gene [15]. If we have p-values then we can define the set of drug affected proteins (DAPs) as follows:

$$DAP = \{v_j | v_j \in V, p_j < 0.05\} \tag{4}$$

B. Measuring the interaction strength

We assumed that the sets of DAPs of the interacting drugs are largely overlapping, which is measured by the Jaccard coefficient. It is 1 if the two sets are identical and 0 if they are disjunct.

$$J(DAP_i, DAP_j) = \frac{|DAP_i \cap DAP_j|}{|DAP_i \cup DAP_j|}, \tag{5}$$

If two drugs share targets or target pathways, then J will be near to 1.

C. Enrichment analysis of interaction causing proteins

For the characterization of the affected proteins we used the concept of Gene Set Enrichment Analysis (GSEA) [16], [17] wherein the goal is to find a common pattern (for instance, most of the genes are parts of the same pathway). The standard approach is based on hypergeometric distribution which has the disadvantage of disregarding the information about how much a given protein is affected by the drug, rather it simply uses the fact that it is affected or not. Instead, we used a modified version of the GSEA which is based on Kolmogorov - Smirnov statistics and predefined function sets. The original method was developed for microarrays where the correlation between the expression of a gene and a phenotype under study is measured [17]. However, the smaller the p-value is the more affected the proteins are, so it can be seen as an *anti correlation* measure, thus it should be converted into a *correlation* like measure r_i. Formally,

$$r_i = \frac{\log(p_i)}{\log(\frac{1}{N_R})} \tag{6}$$

where N_R is the number of randomization ($\frac{1}{N_R}$ is the possible smallest non-zero p-value). The enrichment score of a gene set S (ES(S)) is computed as described in the original paper [17]. $ES(S)$ is large if the set members have low p-values. To assign a statistical significance to the gene set a similar randomization procedure was used, followed by a t-test. In the case of drug combinations $r_i = 0$ if the ith protein is missing from the sets of the DAPs of the components, and p_i was the product of the p-values of the component drugs. The gene sets were downloaded from the official website of Molecular Signatures Database (MSigDB) (retrieved 11-01-2013). For the experiments we used the C2 - CP (canonical pathways) datasets. We also used the gene ontologies (biological process, cellular component, molecular function), which were downloaded from the official site of GO [18] (retrieved 11-01-2013).

D. Description of the experiments

All the algorithms were implemented in MATLAB 2012a. The used network was STRING 9.0 [19] and the drug combination was downloaded from the drug combination database [20]. The drug target data were taken from the STITCH [21] and DrugBank [22] databases.

1) String: STRING (Search Tool for the Retrieval of Interacting Genes) is one of the largest integrated protein interaction databases. The links between the proteins are "associations" (among them several indirect ones) – rather than simple physical interactions. In our experiment we used only these combined interactions and confidences. Only the human proteins and their combined associations were considered.

2) Drug combination database (DCDB): The known drug combination dataset was downloaded from the drug combination database [20] in sql dump file format. Y. Lie et al. [20] classify the drug combinations into two classes - pharmacodynamic and pharmacokinetic interactions - based on the underlying molecular mechanism on action. Pharmacokinetic interactions are those in which one component has an effect on how the other component(s) are absorbed, distributed, metabolized and excreted. Pharmacodynamic interactions are on the other hand those wherein all individual drugs act on the same target or on different targets in the same pathway, on different targets in related pathways, different targets in cross-talking pathways or different targets in pathways of yet unknown relations.

3) Experiments: Since a low number of true negative, unsuccesful DCs is available, we used artificial or random drugs for control purposes. As it was mentioned earlier the drugs' DAP depends on the number of drug targets N_T, the propagation parameter α, and the number of steps taken by the random walker (k). In our experiment we chose $k = 2$ and $\alpha = 0.5$ based on the gene prioritization experience [14]. We used the AUC (area under roc curve) for measuring the ranking performance [23]. For each parameter combination 300 random drugs and their corresponding $DAPs$ were generated and computed. Then the true combinations were compared to 99 random drugs (resampled from the 300 artifical drugs) and an AUC value was optained. For negative control we used random drugs against random drugs (the expected AUC is 0.5). We calculated an average AUC and an average control AUC from hundred repetitions.

III. Results and discussion

Table I shows the average results of the different DC categories. As expected, the randomly generated AUCs are around. It is not surprising that drug pairs affecting the same protein or the same pathway have high AUC values. It is conspicuous however, that some drug pairs having target proteins in unrelated pathways also have high scores (avg. AUC=0.9106), which may point to combinations worthwhile to test experimentally. The high AUC value for the pharmacokinetical interactions is also unexpected, however, the sample size is small (avg number of DCs is 5) so far reaching conclusions can not be made. **Table II** presents the enrichment analysis results of the combination prednisolone (PubChem: CID 5755) and dipyridamole (PubChem: CID 3108) (DC00457 - AUC=0.9942). This drug combination was suggested to have an anti-inflammatory effect by inhibiting inflammatory mediators [24]. One of the most enriched pathways is the CD28 dependent VAV1 pathway that has an important role in the development of T-cells. Other enriched pathways are also related to cell development and the regulation of cell cycle in immune cells such as the ARF pathway or e2f enabled inhibition of pre-replication complex formation.

TABLE I

Identification of experimentally validated drug combinations using neighborhood overlap measures.

Drug combination categories[1]	avg. AUC[2]	avg. r. AUC[3]
Pharmacodynamic categories:		
Different targets of related pathways	0.8764 (137)	0.5003
Different targets of the same pathway	0.9401 (20)	0.5033
Different targets of unrelated pathways	0.9106 (83)	0.4974
Pharmacokinetic categories:		
Same target	0.9907 (17)	0.4948
Distribution	0.8654 (8)	0.5033
Excretion	0.5802 (6)	0.5022
Metabolism	0.8615 (9)	0.5061

[1] Drug combination categories defined in the Drug Combination Database [20].

[2] AUC was calculated from a ranked list of network overlap measures, each list containing one validated combination and 99 random combinations, carried out in 300 repetitions as described in the text.

[3] Random Average Values were calculated in an analogous manner from ranked lists containing one selected random combination and 99 other random combinations. The number of values used to calculate the average is given in parenthesis. Note that the values are close to the theoretically expected value of 0.5.

A. Limitations of the method

The main limitation of the method is that current network databases only report the intensity of the interaction but often do not report the nature of the interactions i.e. whether or not it is synergetic, antagonistic or additive. Neither do they allow one to determine whether the pathways affected by DC components are up or downregulated, or what the actual molecular mechanism is. Because of the scarceness of the available information, the scope of the method will remain limited, and the results can only be regarded as prediction that need to be confirmed by experiment.

IV. Conclusions

In the paper we presented a network based strategy to predict efficient drug combinations based on the hypothesis that pharmacons generate a perturbations that propagate through the network. We assume that drugs that perturb a large number of proteins in common can make efficient combinations, and that the number of jointly perturbed proteins can be estimated by the Jaccard coefficient. A modified gene set enrichment method was used for explaining how the therapeutic effect of the drug combination may emerge since the proteins which are affected by the individual components are known. For testing the hypothesis we used the drug combination database which contains several hundreds of known drug combinations. The method has some limitations, for instance the information included in current protein interaction databases does not allow one to predict whether a drug interaction is synergistic or antagonistic.

TABLE II

List of pathways significantly perturbed by both prednisolone and dipyridamole[1] .

Pathway name[2]	p-value[3]
DCC mediated attractive signaling	0.00023906
SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion	0.00056432
CD28 dependent Vav1 pathway	0.00294221
CDC6 association with the ORC:origin complex	0.00409331
Xenobiotics	0.00416038
Na Cl dependent neurotransmitter transporters	0.00437152
Hyaluronan Metabolism	0.00555056
E2F-enabled inhibition of pre-replication complex formation	0.00572520
Activation of Rac	0.00598630
Linoleic acid metabolism	0.00768036
Transport of organic anions	0.00780610
ARF pathway	0.00783049

[1] The proteins significantly perturbed by both drugs were mapped to the canonical pathways of the Molecular Signatures Database (MSigDB) [17] by gene set enrichment analysis as described in the text. Only those pathways were enriched that have at least 5 but maximum 200 elements [17].

[2] Pathway names are taken from the MSigDB version 3.1.

[3] P values denote the probability for a given pathway occurring at random, calculated as described in the text.

ACKNOWLEDGMENT

BL is PhD student at the Multidisciplinary Doctoral School at the Faculty of Information Technology and Bionics of the Pázmány Péter Catholic University. The work was partially supported by the grants TÁMOP-4.2.1./B-11/2/KMR-2011-0002, TÁMOP-4.2.2./B-10/1-2010-0014 and TÉT 10-1-2011-0058. BGy was supported by the OTKA PD 83154 grant.

REFERENCES

[1] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," Nature Chemical Biology, vol. 4, 682–690, 2008.

[2] K. Hiroaki, "A robustness-based approach to systems-oriented drug design," Nature Reviews Drug Discovery, vol. 5, p. 202–210, 2007.

[3] V. Ágoston, P. Csermely, and S. Pongor, "Multiple weak hits confuse complex systems: A transcriptional regulatory network as an example," Physical Review E, vol. 71, no. 5, http://link.aps.org/doi/10.1103/PhysRevE.71.051909, 2005.

[4] J. Lehár, A. S. Krueger, W. Avery, A. M. Heilbut, L. M. Johansen, E. R. Price, R. J. Rickles, G. F. Short III, J. E. Staunton, X. Jin, M. S. Lee, G. R. Zimmermann, and A. A. Borisy, "Synergestic drug combinations tend to improve therapeutically relevant selectivity," Nature Biotechnology, vol. 27, p. 659–666, 2009.

[5] A-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," Nature Reviews Genetics, vol. 12, no. 1, p. 56–68, 2011.

[6] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," Drug Discovery Today, vol. 12, no. 1-2, p. 34–42, 2006.

[7] A. A. Borisy, P. J. Elliott, N. W. Hurst, M. S. Lee, J. Lehar, E. R. Price, G. Serbedzija, G. R. Zimmermann, M. A. Foley, B. R. Stockwell, and C. T. Keith, "Systematic discovery of multicomponent therapeutics," Proceedings of the National Academy of Sciences, vol. 100, no. 13, p. 7977–82, 2003.

[8] P. K. Wong, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C. M. Ho, "Closed loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," Proceedings of the National Academy of Sciences of the United States of America, vol. 105, no. 13, p. 5105–10, 2008.

[9] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," Molecular Systems Biology, vol. 4:228, Epub 2008 Nov 4.

[10] X. M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and Pharmacological data," PLoS Computational Biology, vol. 7, no. 12:e1002323, Epub 2011 Dec 29.

[11] S. Li, B. Zhang, and N. Zhang, "Network target for screening synergistic drug combinations with application to traditional Chinese medicine," BMC Systems Biology vol. 5, Suppl 1:S10, 2011.

[12] B. Ligeti, Zs. Mihály, Zs. Pénzváltó, R. Vera, B. Győrffy, S. Pongor in preparation.

[13] W. Scott, P. Smyth, "Algorithms for Estimating Relative Importance in Networks," International Conference on Knowledge Discovery and Data Mining -Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, p. 266 –275, 2003.

[14] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau, "Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches," BMC Bioinformatics, vol. 11:460, 2010.

[15] P. N. Westfall, and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley, New York, 1993.

[16] D. W. Huang, B. T. Sherman and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," Nucleic Acids Research, vol. 37, no. 1, p. 1–13, 2009.

[17] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 43, p. 15545–15550, 2005.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology, " Nature Genetics, vol. 25, no. 1, p. 25–29, 2000.

[19] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Research, vol. 39 (Database issue), p. 561–8, 2010.

[20] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: Drug combination database " . Bioinformatics, vol. 26, no. 4, p. 587–588, 2010.

[21] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, LJ. Jensen, P. Bork., "STITCH 3: zooming in on protein-chemical interactions. ," Nucleic Acids Research, vol. 40 (Database issue), p. 876–8, 2012.

[22] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R Eisner, AC. Guo, DS. Wishart., "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.," Nucleic Acids Research, vol. 39 (Database issue), p. 1035–8, 2011.

[23] P. Sonego, A. Kocsor, and S. Pongor, "ROC analysis: applications to the classification of biological sequences and 3D structures," Briefings in Bioinformatics, vol. 9, no. 3, p. 198–209, 2008.

[24] G. R. Zimmermann GR, W. Avery, A. L. Finelli, M. Farwell, C. C. Fraser, A. A. Borisy, "Selective amplification of glucocorticoid anti-inflammatory activity through synergistic multi-target action of a combination drug, " Arthritis research & therapy, vol. 11, issue 1, 2009.