**Method for finding novel drug targets by simulated in-silico gene knockouts.**

**Abstract**

**Introduction**

Precision medicine, an approach where medical treatment is tailored for a specific patient, is now at the forefront of biological and medical research. One of the promising goals of precision medicine is patient-specific drug treatment; it is well known that some drugs affect only a specific sub-group of patients, while even harming other patients with the same disease [1].

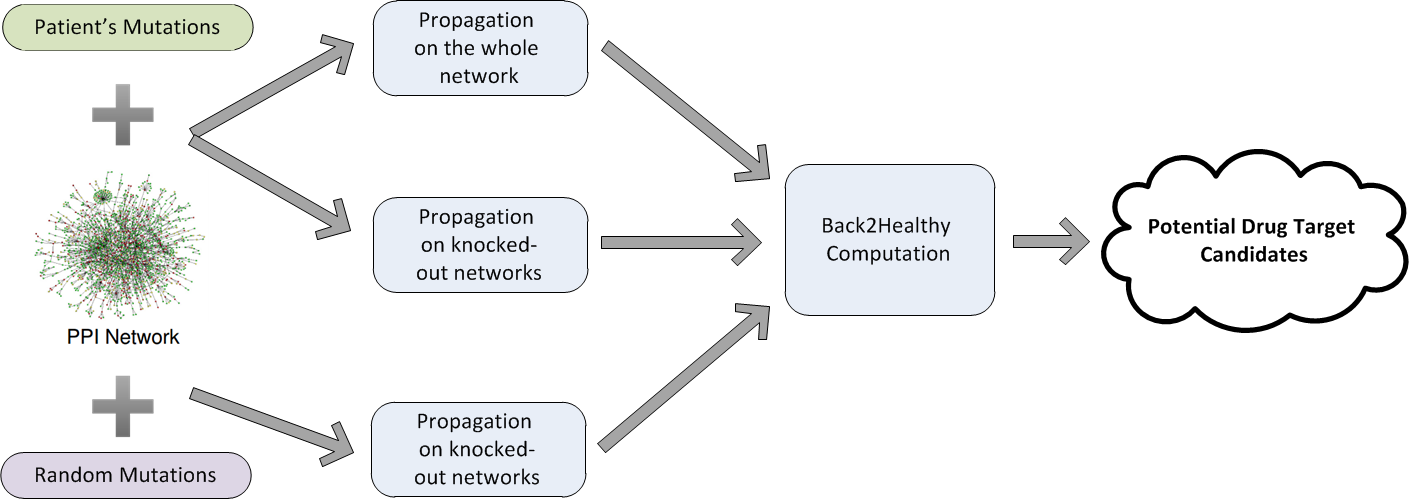
In the past decades, the time and cost of developing new drugs have soared significantly. In general, it takes about 15 years and up to 800 million dollars to convert a promising new compound into a drug in the market [2]. Therefore, many computational approaches arise aiming to identify or predict drug targets in-silico. Those methods tackle the problem in various ways, such as literature text mining [3, 4] and protein three-dimensional structures [5-9]. In particular, the accumulation of various types of omics data gave birth to methods that take advantage of the data to discover novel drug targets [10]. In addition, some methods use protein-protein interaction (PPI) data to infer gene prioritization [11] and to predict novel causal proteins [12]. Li et al. [13] described a workflow that leverages information of network topology to predict novel drug targets using machine learning algorithm.

A common approach in experimental drug discovery is to perform gene knockouts on animal models. Subsequently, methods that use in-silico knockouts on biological derived networks have shown to predict drug targets successfully [14, 15].

The methods mentioned above are all based on non-personalized data; hence they might be ignoring the variance of patients within a specific disease.

In this work we present a novel approach to tackle the drug targets problem from a personalized perspective using in-silico knockouts in a PPI network. As described in figure 1, we start from a general PPI network and personal disease-related data. We rely on the framework described by Vanunu et al. [12] to infer gene prioritization by network propagation. We perform multiple network propagations in order to simulate the current patient state, the patient state after gene knockouts (by removing the gene's node from the network) and an estimated "healthy" state. We use these different states in order to rank the gene knockouts and retrieve a candidates list for potential novel drug targets.

The framework we presented is general and can be applied on any personalized disease-related data. However, to evaluate our performance we applied it on patients suffering from acute myeloid leukemia (AML), using mutated and differentially expressed genes data. We gathered this data from COSMIC cancer gene census [19].



***Figure 1***. The workflow of the proposed algorithmic pipeline. The green node represents input data, the purple node represents generated data, the blue nodes represent processes and the cloud stands for the pipeline's output.

**Results**

We hypothesize that by propagating from individual mutations sets of AML patients we can infer central components in the AML disease. First, we show that we can identify common AML causal genes by synthesizing the individual mutations set propagations and ranking according to propagation scores. Finally, we show that by integrating a smart and personalized knockout process we can infer potential drug targets.

***Calibration***

We ran the algorithm using different parameters and choices for prior knowledge, and tested the resulted sorted gene list for enrichment. We used three sets of known causal genes, varying in confidence and size: 10 AML causal genes from KEGG [ref], 94 AML causal genes from COSMIC and 533 cancer causal genes from COSMIC. It should be noted that for our enrichment tests, we only used the subsets of genes contained within the PPI network, which resulted in 10, 72 and 363 genes, respectively. For more information about the enrichment process, see the Methods section.

The best choice of was estimated to be 0.9. We relied on Vanunu et al. conclusion that any choice of above 0.5 will be sufficient. However we still aimed to refine the algorithm by inspecting three different values above the suggested threshold - 0.5, 0.75 and 0.9. We ran the algorithm on these three values and a sample of 100 patients, receiving a rank for each gene in the network. Aggregating the gene ranks from each patient and receiving a gene list sorted by the propagation score. We checked if this list was enriched using the aforementioned three causal genes sets. All choices of resulted in significant p-values (), though the best one out of the three is 0.9, as can be shown for all three sets of causal genes (Figure 2A). Henceforth, we fixated the parameter to be 0.9.

Next, we went to examine different variants for the prior knowledge gene set P based on patients’ data. Each patient holds information about each gene - whether it is mutated and/or differentially expressed. Therefore we examined four variants. First, setting P to be only mutated genes; Second, setting P to be only differentially expressed genes; Third, performing the first and second options independently and averaging the two propagation scores for each gene; Forth, same as the third option but we take the maximum of the two propagation scores for each gene. Similarly to the comparison between choices of , we tested the enrichment in each of the variants. All prior knowledge variants resulted in significant p-values (). The best variant was setting P to be mutation data only, as can be shown for all three sets of causal genes (Figure 2B). From here on, we use only mutation genes as our prior knowledge set P. This result may be surprising as using only mutations loses components of P that might be due to copy number variations [add ref]. However one must consider that a considerable portion of the genes changed their expression due to the mutations, therefore using the differentially expressed genes in P introduced peripheral noise that was eliminated by using the mutations only.

***Figure 2***. Refining the algorithm's parameters by top 10% enrichment analysis. The red line stands for a p-value of . (A) Comparison between choices of . For KEGG dataset, all choices of resulted in 7 hits out of 10. For COSMIC – AML dataset, choices of 0.5, 0.75 & 0.9 resulted in 23, 23 & 24 hits out of 72, respectively. For COSMIC – cancer dataset, choices of resulted in 112, 121 & 134 hits out of 363, respectively. (B) Comparison between variants for prior knowledge set . . For KEGG dataset, all variants of resulted in 7 hits out of 10. For COSMIC – AML dataset, variants of Maximum, Average, Mutations Only & Expression Only resulted in 26, 26, 28 & 25 hits out of 72, respectively. For COSMIC – cancer dataset, variants of resulted in 142, 145, 148 & 142 hits out of 363, respectively.

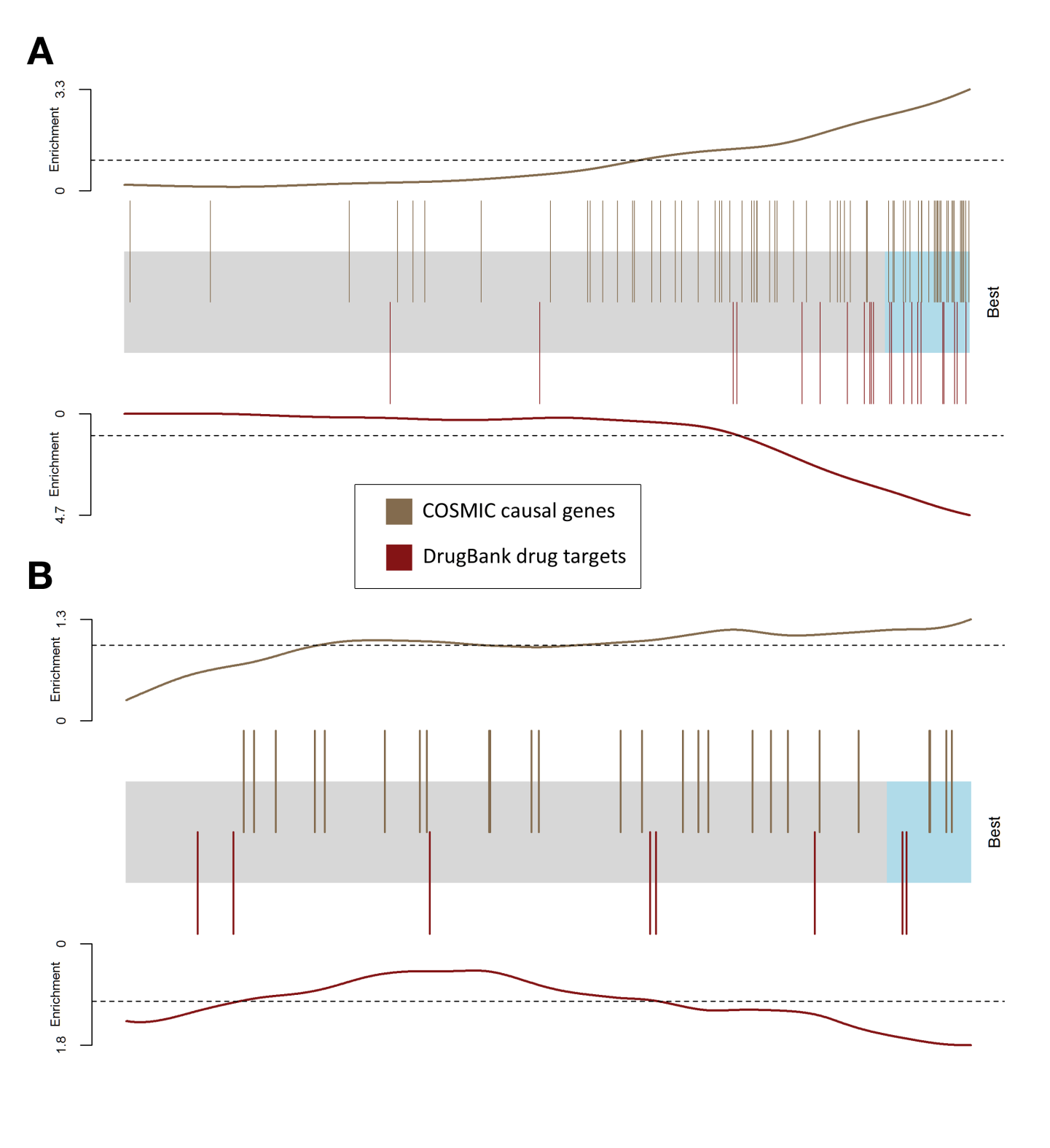
***Revealing potential drug targets***

After refining the propagation algorithm using the discovery of causal genes test, we continue on to knocking out genes for each patient in order to reveal potential personalized drug targets. We do so by eliminating one gene from the graph and running the propagation algorithm, thus receiving a new propagation score for each gene. By comparing these scores to the original scores it can be determined how much did the knockout disrupt the network, or even brought it closer to a healthy state. Our hypothesis is that genes that disrupt the network most are potential drug targets.

In order to measure the disruption caused by the knockout, we chose to use the set of differentially expressed genes of the patient. We chose this set due to two underlying assumptions - differentially expressed genes are the genes that are most affected by the disease [21, 22], and a good drug target would revert the effect of the disease. Hence, a good drug target is expected to significantly revert the effect of the differentially expressed genes.

We took the before and after knockout propagation scores for each of the differentially expressed genes and calculated the Back2Healthy distance score (B2H; See Methods). Then we sorted the knockout genes by the B2H score to retrieve the top disruptive genes as our best candidates for potential personalized drug targets.

This process infers drug targets for each patient. As personalized drug targets do not exist to the best of our knowledge, we turned to validate the aggregated results for all patients, expecting that if our assumption is correct, this list would be enriched with the known AML drug targets. Thus, we gathered two sets of genes to represent AML drug targets, a manually curated list of AML drug targets from DrugBank [18] and a list of causal genes derived from COSMIC database. We created the aggregated list by running the algorithm for all of the patients, receiving a list of potential drug targets for each patient and aggregating it by adding the scores. Thus, naming the top of the joint list to be the best candidates for potential drug targets for the disease. By performing enrichment analysis on the best candidates list using two sets of genes - manually curated drug targets and known causal genes from COSMIC for the disease, we witnessed significant enrichment for both (Figure 3A). To emphasize the importance of the personalized approach we took, we generated a "meta-patient", using consensus mutations and differentially expressed genes derived by aggregating all AML patients. Consensus genes were defined as genes appearing in at least 5 patients' data. Running the pipeline on the meta-patient resulted in insignificant enrichment (Figure 3B), therefore stressing the advantage of a personalized approach.



***Figure 3***. Aggregated drug target candidates' barcode plot. The knocked out genes are represented by a shaded rectangle, where the top 10% (drug target candidate list) are shaded cyan. The colors differ between collections of gene sets that are known or expected to be good drug targets, where every overlaid bar stands for a single gene in a collection. Traces above/below the bar represent relative enrichment. (A) The barcode plot was generated by running our method on each AML patient independently and aggregating the results. (B) The barcode plot was generated by running a similar single pipeline on the meta-patient.

**Materials and Methods**

***Computing propagation score***

We use the network propagation method described in Vanunu et al. [12]. The input consists of a network with V as the set of proteins, E as the set of their interactions, represents the reliability of the interaction between and , and a prior knowledge protein set . Our goal is to prioritize the proteins in with respect to . We do so by defining a function that is both smooth over the network and accounts for the prior knowledge about each node.

Regarding our framework, the set of protein-protein interactions and their reliability are taken from the HIPPIE network [16]. The prior knowledge set is derived from the patient's mutation data.

As described by Vanunu et al. we use Laplacian normalization to produce the normalized network edge weight . Briefly we construct a sparse matrix from the edge weights , and construct a diagonal matrix with . The normalized edge weight matrix is computed as .

We define a prior knowledge function such that:

With the normalized weight matrix and the prior knowledge function , we use the iterative procedure described by Zhou et al. [20] to compute . Namely, starting with , we update at iteration as follows:

The procedure is repeated iteratively until convergence; namely we stop the iterations when the following condition occurs:

The propagation score for each gene is its rank among all genes after propagating the network, where lower ranks means higher value. In case of ties the ranks of the genes is averaged.

***Enrichment Analysis***

Throughout our work, we used enrichment analysis to test a hypothesis where a set of interesting genes is more highly likely to be in the top of the sorted list of genes.

Let be the size of the sorted list of genes, be the size of the set of interesting genes (), be the number of interesting genes that appear in the top genes in the sorted list. We performed a hypergeometric test, where we check the probability to witness or more successes in draws from a finite population of size that contains exactly successes. To define the top of the list, we defined to be .

***Back2Healthy distance score***

Let , be vectors of propagation scores for a gene set . was generated by propagating on the original PPI network, while was generated by propagating after the removal of a knockout gene. We define the Back2Healthy (B2H) distance between and as follows:

First, define .

For , we generate a score vector for by propagating the original PPI network and setting the prior knowledge set to be random nodes. Thus, simulating a “healthy” distribution of propagation scores for .

Next, for , define

And

Hence, represents the quantile of in our simulated distribution, same goes for . For our method, we used .

Finally, is defined by:

**Discussion**

**References**

1. Rothwell, Peter M. "Can overall results of clinical trials be applied to all patients?." *The lancet* 345.8965 (1995): 1616-1619.
2. J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, “The price of innovation: new estimates of drug development costs,” Journal of Health Economics, vol. 22, no. 2, pp. 151–185, 2003.
3. S. Zhu, Y. Okuno, G. Tsujimoto, H. Mamitsuka, A probabilistic model for mining implicit chemical compound-gene relations from literature, Bioinformatics 21 (2005) ii245–ii251.
4. R.B. Altman, C.M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L.J. Jensen, M. Krallinger, B. Mons, S.I. O'Donoghue, M.C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay, A. Valencia, Text mining for biology-the way forward: opinions from leading scientists, Genome Biol. 9 (2008) S7.
5. A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, M. Rarey, Combining global and local measure for structure-based druggability predictions, J. Chem. Inf. Model. 52 (2012) 360–372.
6. E. Perola, L. Herman, J. Weiss, Development of a rule-based method for the assessment of protein druggability, J. Chem. Inf. Model. 52 (2012) 1027–1038.
7. P. Schmidtke, X. Barril, Understanding and predicting durggability. A highthroughput method for detection of drug binding sites, J. Med. Chem. 53 (2010) 5858–5867.
8. A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts smallmolecule druggability, Nat. Biotechnol. 25 (2007) 71–75.
9. A. Krasowski, D. Muthas, A. Sarkar, S. Schmitt, R. Brenk, DrugPred: a structurebased approach to predict protein druggability developed using an extensive nonredundant data set, J. Chem. Inf. Model. 51 (2011) 2829–2842.
10. H. Tan, X. Ge, and L. Xie, “Structural systems pharmacology: a new frontier in discovering novel drug targets,” Current Drug Targets, vol. 14, no. 9, pp. 952–958, 2013.
11. S. Erten, G. Bebek, and M. Koyut¨urk. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. Journal of computational biology, 18(11):1561–1574, Nov. 2011.
12. O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*., 6(1):e1000641, Jan 2010.
13. Li, Zhan-Chao, et al. "Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network."*Analytica chimica acta* 871 (2015): 18-27.
14. Ciaccio, Mark F., et al. "The DIONESUS algorithm provides scalable and accurate reconstruction of dynamic phosphoproteomic networks to reveal new drug targets." *Integrative Biology*, 2015.
15. Fatumo, Segun, et al. "Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico."*Infection, Genetics and Evolution* 9.3 (2009): 351-358.
16. M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2):e31826, 02 2012.
17. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328, 2004.
18. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z.T. Dame, B. Han, Y.
19. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Research, 39(suppl 1):D945–D950, 2011.
20. Zhou, D.S. Wishart, DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. 42 (2014) D1091–D1097.
21. Van't Veer, Laura J., et al. "Gene expression profiling predicts clinical outcome of breast cancer." *nature* 415.6871 (2002): 530-536.
22. Valk, Peter JM, et al. "Prognostically useful gene-expression profiles in acute myeloid leukemia." *New England Journal of Medicine* 350.16 (2004): 1617-1628.