FPN: 8P

A Gradient Boosting Decision Tree (GBDT) Approach to Identify Potential Therapeutic Targets

Ilona Kifer, Elinor Dehan, Eden Goldfarb, Shay Rotkopf, Gabi Tarcic and Michael Vidne

FORE Biotherapeutics, Ness Ziona, Israel

Introduction

Identification of novel therapeutic targets and associated biomarkers is the first step in the drug development journey, and traditionally relies on deep understanding of the underlying biology, a process which is lengthy and non-scalable. The advent of high-throughput screening methods and advanced machine learning (ML) tools enable rapidly uncovering novel therapeutic targets and biomarkers even where the biology is not yet well understood.

Here we present the use of a supervised tree-based approach to predict cell line-drug sensitivity or resistance and to estimate which genomic features influence these the most. These features, specific mutated genes and combinations thereof, represent potential therapeutic targets of each screened drug. We evaluated our method's performance on 8 drugs with well-established putative genomic targets and then employ the method on the entire set of CancerRx dataset. For the evaluation set we show that in addition to the known target, other genomic features were identified for most of the drugs evaluated, including some not previously reported. These may represent additional novel targets or biomarkers of sensitivity or resistance, to potentially extend benefit to more patients or to better select patients most likely to benefit, respectively.

Data And Methods

Data acquisition: Cell-line genomic profile data from Sanger institute's Cell Model Passports website was represented as a matrix of 1357 cell lines vs. 298 cancer-related genes, indicating the mutational status (WT/MT) of the gene in each cell-line. Cell-line drug sensitivity data of IC50 values transformed into z-scores, was acquired from the CancerRX GDSC1 dataset of 945 cell-lines tested across 345 compounds.

Model: Our ML approach was trained to predict the sensitivity of all cell-lines to each drug. We used the XGBoost package to implement a Gradient Boosting Decision Tree (GBDT) machine learning algorithm.

Training & validation: The input to our model is the cell-lines genomic profile matrix and the response of each cell-line to the given drug. We randomly selected 80% of the cell-lines to train the model and used the remaining 20% for model validation. We repeated sample selection and model training 200 times to correct for a bias caused by the scarcity of sensitive cell lines.

Feature importance: To estimate the effect of each genomic feature on the predicted drug response we use SHAP values (denoted as $\boldsymbol{\psi}$), which are calculated per cell-line and genomic feature and measure the difference in prediction between a model trained with the feature and in its absence. A negative/positive $\boldsymbol{\psi}$ value indicates a feature that contributes to the sensitivity/resistance of the cell-line to the drug, respectively. The further from 0 the $\boldsymbol{\psi}$ value is, the larger the feature's effect on the model's prediction.

Detecting potential biomarkers: A potential biomarker is a gene that shows a large differential contribution between its mutated and WT state for a given drug. Thus, for each genetic feature we compute two values:: $\boldsymbol{\psi}_{\mathrm{MT}} = mean(\boldsymbol{\psi}(mutated\ cell\ lines))$ and $\boldsymbol{\psi}_{\mathrm{WT}} = mean(\boldsymbol{\psi}(WT\ cell\ lines))$. Our potential biomarker score is calculated for every gene as:

$$\boldsymbol{\psi}_{\mathrm{diff}} = \boldsymbol{\psi}_{MT} - \boldsymbol{\psi}_{\mathrm{WT}}$$

Methodology Outline

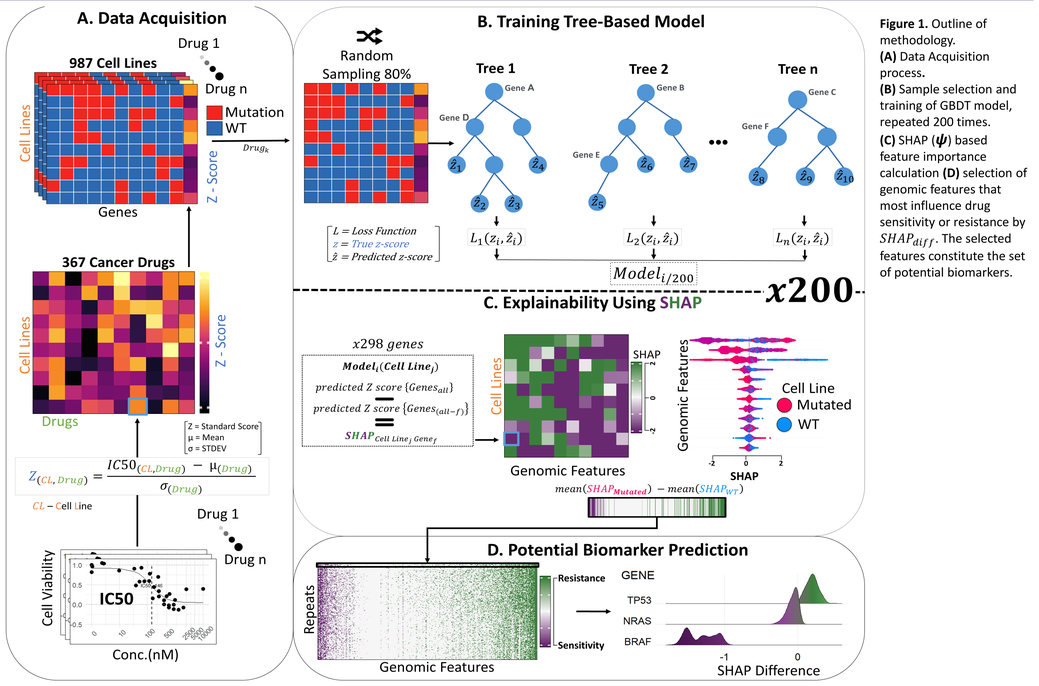

Figure 1. Outline of methodology. **(A)** Data Acquisition process. **(B)** Sample selection and training of GBDT model, repeated 200 times. **(C)** SHAP ($\boldsymbol{\psi}$) based feature importance calculation **(D)** selection of genomic features that most influence drug sensitivity or resistance by $SHAP_{diff}$. The selected features constitute the set of potential biomarkers.

Results on Validation Set

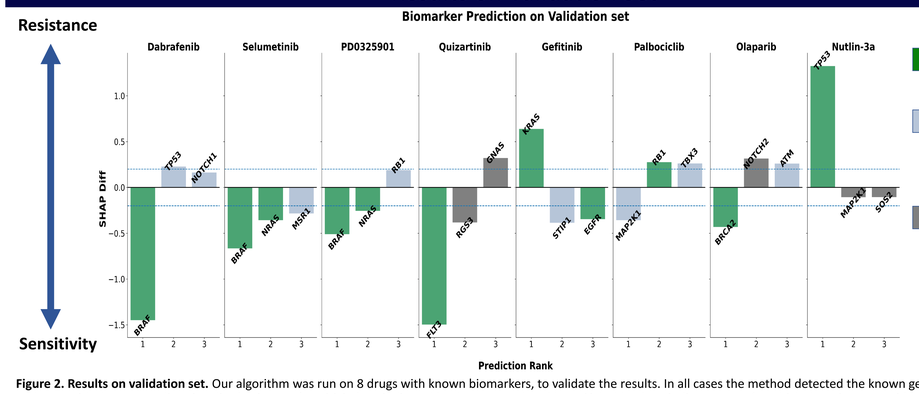

Figure 2. Results on validation set. Our algorithm was run on 8 drugs with known biomarkers, to validate the results. In all cases the method detected the known genomic feature(s) or pathway, with the correct directionality (resistance/sensitivity), as indicated by the green colored bars. Light blue bars indicate genes for which there is some evidence in the literature, and gray bars indicate novel findings of the algorithm for which there is little to no other evidence in the literature. Shown are the top 3 genes for each drug. The biomarkers with the highest absolute $\boldsymbol{\psi}_{\mathrm{diff}}$ for each drug are usually direct targets of the relevant drug. These are followed by additional biomarkers that can often be linked to the pathway or have other supporting evidence for the response we observe. **Dabrafenib:** BRAF is the direct target, while TP53 and Notch1 mutations were shown to lead to BRAF/MAPK inhibitor resistance. **Selumetinib:** the MAPK pathway is its direct target and also related to increased MSR1 expression, explaining MSR1 sensitivity to the drug. **PD0325901:** a MEK inhibitor, MAPK being its target pathway. RB1 mutation is associated with resistance to the drug. **Quizartinib:** FLT3 is its direct target. **Gefitinib:** EGFR is its direct target and KRAS is known to cause resistance to the drug. STIP1 is a co-chaperone with Hsp90, which is important for the stability and activity of EGFR, consistent with STIP1 sensitivity to Gefitinib. **Palbociclib** is a CDK inhibitor. The MAPK pathway is known to enhance CDK4/6 activity, in-line with Palbociclib sensitivity in a MAP2K1 mutant background. RB1 mutations are also associated with resistance to Palbociclib. Additionally, It was shown that TBX3 deficiency leads to decreased CDK4, which can explain TBX3-related resistance to Palbociclib. **Olaparib:** BRCA2 is the direct target. ATM has been known to be involved in DNA damage repair and considered as a potential target of PARP inhibitors. **Nutlin-3a** target is MDM2 which binds TP53 and causes its degradation. It does not affect a mutated TP53. Thus, TP53 is a biomarker for drug resistance.

Large-Scale Exploration

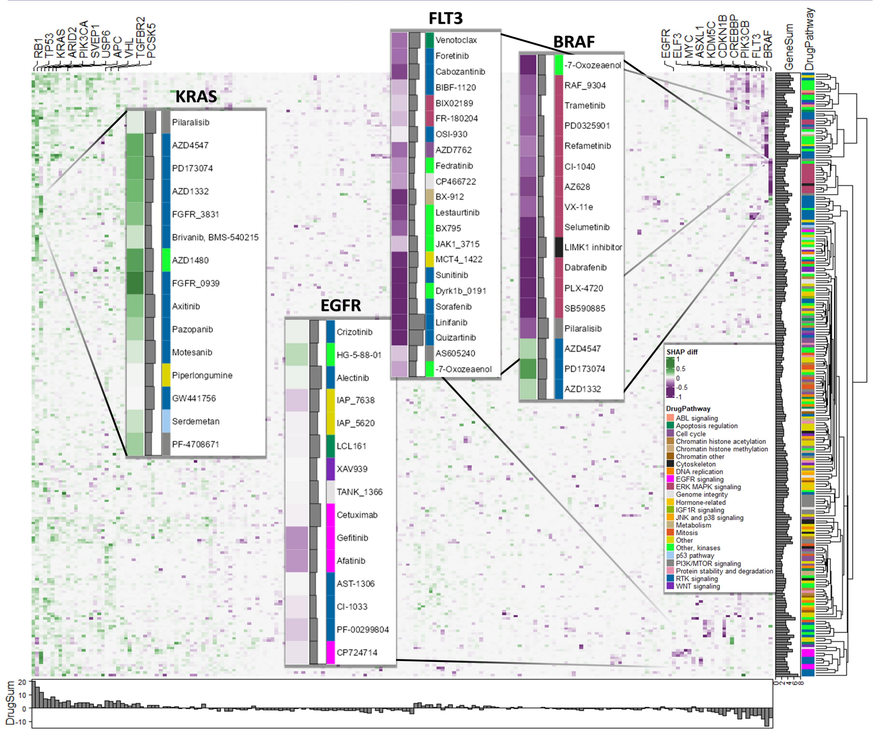

Figure 3. To explore the space of potential drug-target combinations, we ran the algorithm on the set of 346 drugs present in the CancerRx GDSC1 dataset. The heatmap presents the $\boldsymbol{\psi}_{\mathrm{diff}}$ value of each genomic feature (columns) for every drug (rows, color-coded according to drug pathway). Rows and columns were hierarchically clustered by $\boldsymbol{\psi}_{\mathrm{diff}}$ distance similarity using the Bray-Curtis metric . Also shown are the column-wise sum (for each gene, summing all drugs), highlighting the most sensitive and resistant biomarkers. The zoom-ins indicate groups of drugs for which one or more genomic feature were found to correlate with drug sensitivity (purple) or resistance (green).

Conclusions

Here we present a novel approach to explore the space of potential drug targets and biomarkers. We validate our approach on a set of 8 drugs with known targets and biomarkers for drug sensitivity and/or resistance and show our ability to detect them by calculating $\boldsymbol{\psi}_{\mathrm{diff}}$ values per drug and genomic feature. We also identify additional genomic features, some previously unreported.

A large-scale application of this approach yields many additional findings, some of which were known, and some require further investigation. An interesting observation is that while many drugs such as MAPK inhibitors have a similar target, they still have a distinct sensitivity profile to the set of 300 genes.

We note that the presented methodology is an initial proof of concept. We plan to further develop it in several directions: first, by expanding our genomic feature set to differentiate between different mutation types, e.g., mutation classes, fusions, CNAs and other onco-relevant aberrations. Second, We aim to explore alternative definitions to genomic features, e.g., genetic families, pathways or other connections. We believe our approach has the potential to highlight non-trivial drug-gene relationships, to aid in the detection of novel therapeutic biomarkers for cancer, and thus to provide indications and guidance towards a better selection of patients for specific drug treatment.

Contact Information

Corresponding authors can be contacted at Ilona.kifer@fore.bio, gabi.tarcic@fore.bio, or michael.vidne@fore.bio. Authors declare no conflict of interests.