Machine Learning Approaches for Drug Response Predictive Models

RCDML Project: Mauricio H. Ferrato[1], Adam Marsh[1, 3], Karl Franke[2], Sunita Chandrasekaran[1], Erin L. Crowgey[2]
Candle Project: Vineeth Gutta[1], Satish Ranganathan[4], Matthew Beyers[4], Sara Jones[4], Eric Stahlberg[4], Sunita Chandrasekaran[1]
Institution Affiliations - University of Delaware[1]; Nemours Children Health System, Wilmington Delaware[2]; Genome Profiling LLC[3]; NCI/NIH/FNL[4]

Introduction

Precision medicine efforts have focused on utilizing an individual's genomic profile to aid in the clinical decision-making process.

Application of machine learning techniques is this field have demonstrated both successes and challenges.

The **RCDML project** consists of an ML framework that can:

- Classify a subject as a potential responder versus a non-responder to a given therapy accurately
- Help explain and unravel some of the genomic complexity of the disease.
- Provide a framework that is robust, specific, and confident at classifying a responder

Improving **CANDLE**'s cell line based deep learning (DL) models:

- Test efficacy of the variou DL models on the new data
- Incorporate patient data to predict drug response for individuals
- Identify best algorithms for specific cancers

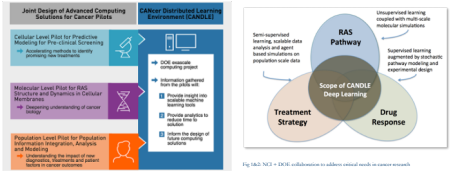

Fig 1&2: NCI + DOE collaboration to address critical needs in cancer research

Methods

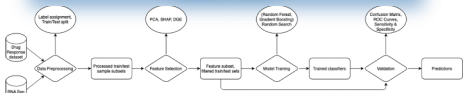

Figure 2. **R**NA-seq **C**ount **D**rug Response **M**achine **L**earning (**RCDML**) framework

Collectively, the RCDML workflow consist of the following steps:

- **Data Preprocessing:** RNA-seq count data and drug response data gets loaded. Assigning samples to high/low responder group and splitting dataset in folds
- **Feature Selection:** Principal Component Analysis (PCA), SHAPLEY, and Differential Gene Expression (DGE) analysis used to reduce RNA-seq count data matrix and identify top candidates for possible biologically insightful features
- **Classifier Training:** Model training using Gradient Boosting (xgboost) and Random Forest (scikit-learn). Hyperparameter optimization by randomized grid search cross-validation
- **Validation:** Sensitivity vs. specificity is analyzed using confusion matrix from fold predictions and receiver operating curves (ROC) are used to compare for every feature selection + classifier combination tested

Experiments & Analysis

RCDML Experimental Setup:

- To test the *RCDML* framework we used the Beat AML[1] dataset
- The Beat AML dataset consist of:
  - *the RNA-seq count data* - count per million (cpm) counts of 22,844 genes for 451 clinical pat
  - *Ex vivo drug response data* - ic50 and AUC scores for 122 different drugs/inhibitors
  - Results focus on 24 drugs that are members of the *RTK-Type-III* inhibitor family
- Clinical patients are assigned to high responder/low responder group
  - *High responder* - drug response AUC score is > 0.75 quantile
  - *Low responder* - drug response AUC score is < 0.25 quantile
- 30 features were selected for each CV fold using a feature selection
- Conserved features were identified across fold iteration
- All of the RCDML experiments were run using the DARWIN machine

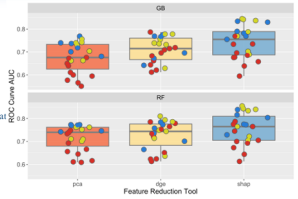

Figure 3. The distribution of ROC-AUC scores for the 24 drugs models. Each whisker plot is a feature selection + classifier combination. The scores are color coded based on the models' feature conservation index.

RCDML Analysis:

Feature selection analysis (Figure 3): The SHAP feature selection technique (purple box-whisker) out-performed PCA and DGE (ANOVA $p < 0.0001$ and Tukey HSD, $p < 0.0001$) in regards to AUC for the ROC curves

Conserved feature analysis (Figure 3): The number of drug AUC scores with conserved feature index in the top bin ">0.5" that were greater than the median AUC for each feature reduction + classifier contrast was significantly higher than the ">0.5" count that were less than the median.

Classifier analysis (Figure 4): No statistical difference in performance was found on the impact of the overall classification performance by using either classifier.

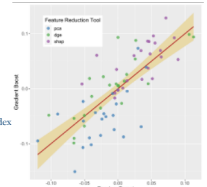

Figure 4. Gradient boosting vs the random forest auc scores. The slope is not sigdiff from 1 and the intercept is not sigdiff from 0.

Extending **CANDLE**'s capabilities:

- CANDLE Pilot 2 includes models such as Combo[2] (Fig. 5)
- Test if models trained with cell line based data such as Combo can be applied to tumor-normal tissue data.
- DARWIN cluster was integral for this work

Methodology

- Obtain tissue data from public databases such as GTEX and GDC
- Obtain drug response data from PharmacoDB
- Format the data according to the ML method requirement.
- Run the ML algorithms and optimize the parameters using CANDLE
- Evaluate and compare performance vs existing cell line based models

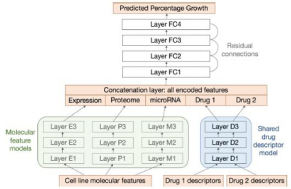

Fig 5. Neural net architecture. The orange boxes, from bottom to top, represent input features, encoded features, and output growth values. Green box indicates molecular features and blue indicates drug features. All encoded features are combined to form input for the fully connected layers at the top. Image is from the Combo paper referenced.

Discussion & Conclusion

Results on the left, show that feature selection technique, rather than the classifier, had the greatest impact on model performance with – the SHAPLEY technique outperforming the other feature selection techniques with an ROC-AUC mean score of .76.

Large source of variance was the large difference in response rates (AUC) from different drug inhibitors since each drug inhibitor is different.

More analysis needs to be done on the specifics of each drug model and the biological impact related to the genes selected by the feature selection techniques.

Future Work

- For RCDML model:
  - Pathway analysis is being worked on the genes selected by feature selection to demonstrate that at the time of diagnosis a trancriptomics signature exists that can predict response to treatment
  - Confidence and robustness experiments on each drug model by checking impact of model performance by changing the feature signature
- Initial experimentation and analysis for CANDLE was focused on the Combo model. Other models this work must be extended to include Uno, UnoMT, and P1B3

References & ACK

1. Burd, A., Levine, R.L., Ruppert, A.S. *et al.* Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. *Nat Med* 26, 1852–1858 (2020). https://doi.org/10.1038/s41591-020-1089-8
2. Xia, F., Shukla, M., Brettin, T. *et al.* Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* 19, 486 (2018). https://doi.org/10.1186/s12859-018-2509-3

Acknowledgements:

- The Lisa Dean Moseley Foundation (PI Crowgey) for funding this project and the Nemours Biomedical Research Department for supporting the RCDML work
- National Cancer Institute (NCI) at the Frederick National Lab for cancer research (FNL) and Leidos for supporting the research on CANDLE
- The DARWIN computing project at the University of Delaware by NSF award OAC-1919839