Predicting Drug Response of Cell Lines from Gene Expression Data

Jacob Kelly[1,2], Arvind Mer[1], Sisira Nair[1], Hassan Mahmoud[1], Benjamin Haibe-Kains[1,2,3]

[1]Princess Margaret Cancer Centre, University Health Network; [2]Department of Computer Science, University of Toronto; [3]Vector Institute for Artificial Intelligence

ABSTRACT

The promise of precision medicine is to individualize cancer care for patients via discovery of biomarkers in molecular profiling. Towards this aim, we investigate the presence of biomarkers in gene expression data for acute myeloid leukemia (AML). Specifically, we develop drug-specific machine learning models for predicting the area under the drug-dose response curve (AUC) of cell lines based on their gene expression. We apply our model to two pharmacogenomic datasets, one consisting of data from immortalized cell lines, the other ex-vivo primary cell lines from patients. We train and cross-validate on immortalized cell lines and examine the generalization of our model to ex-vivo cell lines. We use linear regression with elastic net regularization as our model and compare methods for feature selection.

PROBLEM FORMULATION

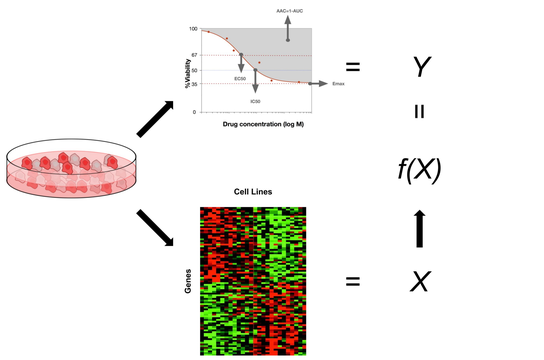


Figure 1. We formulate the problem as a machine learning problem. For each drug, we train a model to predict the AUC of a cell line for the drug based on its gene expression.

DATASETS

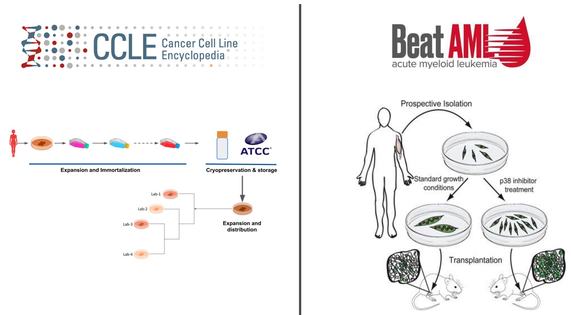


Figure 2. The Cancer Cell Line Encyclopedia (CCLE)[1] dataset consists of immortalized cell lines. The Beat AML dataset consists of ex-vivo patient derived cell lines.

LINEAR REGRESSION

$$n = \text{number of examples}$$
$$m = \text{number of features}$$

$$y^{(i)} = \sum_{i=1}^{m} w_j x_j^{(i)} + b$$

Equation 1. The predictions of the model are an affine linear function of the input features.

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - t^{(i)})^2 + \lambda \mathcal{R}$$

Equation 2. The loss used to train the model; the average squared error of the model's predictions and a weighted regularization term, explained in the following section.

ELASTIC NET REGULARIZATION

$$\mathcal{R} = \alpha \sum_{j=1}^{m} |w_j| + (1 - \alpha) \sum_{j=1}^{m} w_j^2$$

Equation 3. Elastic net regularization is a convex combination of the penalties used for lasso and ridge regularization.

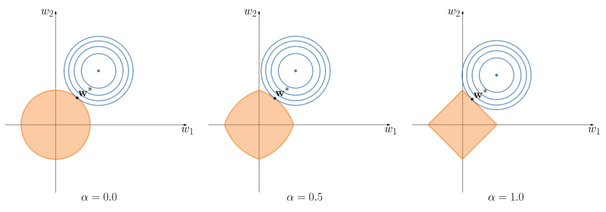


Figure 3. An illustration of the geometry of the elastic net regularization term for different convex combinations. At either extreme, elastic net simplifies to the ridge and lasso terms, respectively. Ridge is known to uniformly encourage small weights, while lasso is known to encourage sparse weights, as illustrated. Elastic net's interpolation combines these two properties.

FEATURE SELECTION

Pathway	Select genes with greatest transcript similarity coefficient[2] (TSC), calculated between CCLE and Beat AML datasets.
Variance	Select genes with greatest variance over all cell lines in the training set.
Univariate	Select genes with largest correlation with target over all cell lines in the training set.
Random	Randomly select genes without replacement.

Table 1. We compare four methods for selecting which features to train our model on. We reduce the number of features from ~19 000 (protein-coding) genes to ~5000 genes.

PIPELINE

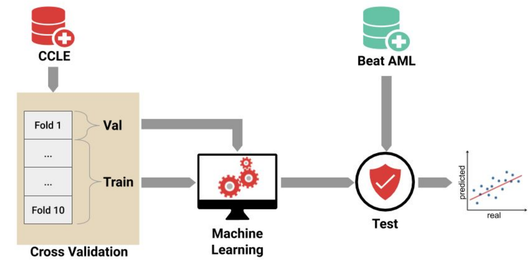


Figure 4. Our model is trained and its hyperparameters optimized via 10-fold cross-validation on CCLE. We test our model for its generalization to unseen test data from Beat AML.

RESULTS

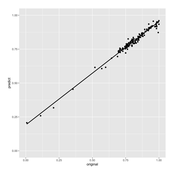
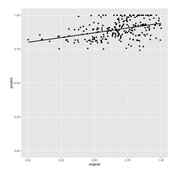
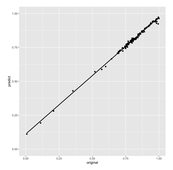
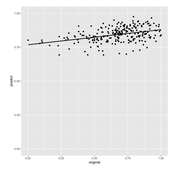
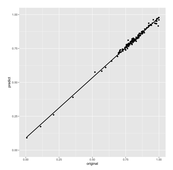
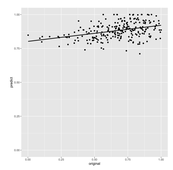
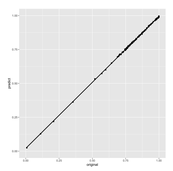
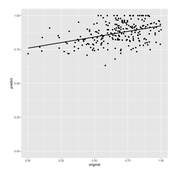
	CCLE		Beat AML		
Pathway		Correlation: 0.989 (p<0.0001)		Correlation: 0.401 (p<0.0001)	α=0.0 λ=1.0
Variance		Correlation: 0.998 (p<0.0001)		Correlation: 0.393 (p<0.0001)	α=0.0 λ=0.0
Univariate		Correlation: 0.996 (p<0.0001)		Correlation: 0.376 (p<0.0001)	α=0.0 λ=0.0
Random		Correlation: 1.000 (p<0.0001)		Correlation: 0.402 (p<0.0001)	α=0.2 λ=0.0

Figure 5. Model prediction results. Optimized hyperparameters reveal that model often degenerates to ridge regression. Model appears to suffer from severe overfitting.

Acknowledgements

Pattern Recognition and Machine Learning, Christopher Bishop. (Inspiration for Figure 3)
Ian Smith, Petr Smirnov, and Zhaleh Safikhani. (Figures 1 & 2)
Arvind Mer. (Figure 4)
BHK Lab Members

References

[1]Barretina, Jordi et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." *Nature.* (2012)
[2]Madani Tonekaboni, Seyed Ali et al. "SIGN: similarity identification in gene expression." *Bioinformatics.* (2019)