TARGET COMPOUND IDENTIFICATION

William Sheerin
william.sheerin@emc.com

Wei Lin
wei.lin@emc.com

Ronaldo Braga
ronaldo.braga@emc.com

Shriya Chintada
sai.chintada@emc.com
Table of Contents
Introduction ................................................................. 3
Objectives ........................................................................ 3
Data and Methodology ......................................................... 4
  Data Cleaning .................................................................. 5
  Descriptive Analysis .......................................................... 5
Results ............................................................................... 9
Discussion of Results ........................................................... 16

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell EMC’s views, processes or methodologies.
Introduction
Target Compound Identification is the basis of drug discovery and design. It is time-consuming and the development costs can be as high as one billion dollars. Within the life sciences industry, it takes 10-17 years to develop an effective treatment. There is a need to identify potential compounds that exhibit beneficial change (efficacy) early in the drug research process. Once identified, the research groups need to capture the effectiveness of targets during clinical trials that have minimal side effects (adverse events). One of the most challenging research aspects is to understand how the target human cells (healthy, benign and malignant) react to oncology drugs and environment changes. Another important goal is to determine the effectiveness of the drug by analyzing the dose response of an effect target. Using these kinds of approaches, it is possible to predict which drugs will better perform in human clinical trials.

Ideally, by utilizing this analysis, the researchers will be able to streamline the development process to effectively identify compounds that are suited for human trials. The sooner the effectiveness of the compounds can be identified, the greater the operational efficiency of the research will be. Moreover, through the utilization of this analysis within a collaborative platform, companies can hope to see ROI of over 500%. How will this be achieved? First, by sharing this clinical information with other clinical teams via a real-time dashboard redundant or unnecessary testing will be eliminated. Additionally, as effective compounds are identified the clinical teams can quickly collaborate on identifying the appropriate dose response. Even incremental efficiencies gained in this manner can reduce operational costs and aid in eventually bringing effective therapies to the marketplace.

Objectives
Using a real dataset describing a set of drug compounds, the goal of the present work is to develop a framework that will identify drug compounds that have a greater go-to-market potential. The objectives can be summarized as follows.

- Target Validation of potentially active compounds modulators
- Performing biomarker identification
- Determine effectiveness by coupling with high-resolution images
Data and Methodology

After analyzing various datasets, the team chose the dataset “Screen20120_Fallahisichani.xls online on “http://lincs.hms.harvard.edu/data/bundles/” to develop the analytics framework. The analysis was developed in R (Open Source Analytics Tool), following the steps mentioned in the paper “Metrics other than potency reveal systematic variation in responses to cancer drugs” (Fallahi-Sichani et al, 2013).

The analysis performed with the referred dataset used regression analysis to fit a model with coefficients to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.

Evaluating the data in the exploratory analysis it was possible to understand that the number of rows to be analyzed individually would be insufficient to obtain an outcome. Thus, the team decided to go through a clustering process in order to identify similar groups to be modeled.

The following is the abstract from Fallahi-Sichani et al. (2013) Nature Chemical Biology: Large-scale analysis of cellular response to anticancer drugs typically focuses on variation in potency (half-maximum inhibitory concentration, (IC50)), assuming that it is the most important difference between effective and ineffective drugs or sensitive and resistant cells. We took a multi-parametric approach involving analysis of the slope of the dose-response curve, the area under the curve and the maximum effect (E_max). We found that some of these parameters vary systematically with cell line and others with drug class. For cell-cycle inhibitors, E_max often but not always correlated with cell proliferation rate. For drugs targeting the Akt/PI3K/mTOR pathway, dose-response curves were unusually shallow. Classical pharmacology has no ready explanation for this phenomenon, but single-cell analysis showed that it correlated with significant and heritable cell-to-cell variability in the extent of target inhibition. We conclude that parameters other than potency should be considered in the comparative analysis of drug response, particularly at clinically relevant concentrations near and above the IC50.

The variables in the dataset include; - “Small Mol HMS LINCS Batch ID”, “Small Mol Name”, “Cell Name”, “maxDose(log10)”, “R^2” (goodness of fit), “EC50(log10)”, “IC50(log10)”, “GI50(log10)”, “hillslope”, “E_inf”, “E_max”, “areaUnderCurve”, “Recorded Plate”, “Recorded Well”, “Control Type”.

With reference to the preliminary model we developed, the target variable used was “IC50 (log10)”. IC50 is a term used in a number of ways: (i) the molar concentration of an antagonist that reduces the response to an agonist by 50%; the concentration of agonist should be given; (ii) the molar concentration of an unlabeled agonist or antagonist that inhibits the binding of a radio ligand by 50%; the concentration of radio ligand should be given; (iii) the molar concentration of an inhibitory agonist that reduces a response by 50% of the maximal inhibition that can be attained.

The nine dose-response metrics which are all the numerical variables used in our analysis are maxDose (log10), R^2 (goodness of fit), EC50 (log10), IC50 (log10), GI50 (log10), hillSlope, E_inf, E_max, areaUnderCurve.
Data Cleaning

The raw dataset had a number of missing values and columns such as Recorded Well, Control Type had null values. The data cleaning process included eliminating the null and missing values in the dataset.

Descriptive Analysis

The graph below is an example to explore the dose-response data analyzed in Fallahi-Sichani, et al. (2013), by overlaying multiple "cell line vs. cell line" or "drug vs. drug" scatterplots. The different scatterplots shown in the right-hand panel correspond to different dose-response metrics (IC50 (log10), GI50 (log10), HillSlope (log2), E_max).

Here, we have selected the “drug vs. drug” plot for two different combinations (blue and pink). The blue pair represents the drug ‘Baicalein’ on the x-axis and ‘Temsirolimus’ on the y-axis. The pink pair represents the drug ‘Baicalein’ on the x-axis and ‘Ixabepilone’ on the y-axis.

Figure 1: Observation vs. Values which shows missing values

Figure 2: Scatterplot for all four dose-response metrics based on drugs
Another analysis involved determining the highest occurrence of a Small Molecule-based on the count of the cell names presents within the data which is represented by the tree graph below.

‘MG-132’ small molecule evidently had the highest occurrence with a count of 106 cell names associated with it.

![Small Molecule Names](image)

**Figure 3: Small Molecule Names where the color and count show the ‘count’ of Cell name**
‘MG-132’ small molecule was then considered to see the relationship between EC50 and IC50 values. It’s observed that for this small molecule, IC50 and EC50 values are seen highest for the cell name ‘BT-483’. A little bit of background around this particular cell name BT-483 says that it’s a cell belonging to a female Homo sapien, and the organ being breast and the disease detail indicated is “breast ductal carcinoma”.

![Figure 4: Small Molecule](image)

Another interesting box-chart analysis (Figure 5) showed that ‘5-FU’ and ‘Valproic acid’ small molecules are given to cells in high doses whereas MG-132 small molecule is given in low dose. The median range being -4

![Figure 5: Cell Name vs. maxDose (log10) marked by small molecules](image)
For the nine dose-response metrics, we plotted out the density for each metric to understand the distribution of data and the skewness. Clearly, $E_{\text{max}}$ and $E_{\text{inf}}$ metrics have similar distribution with a peak at 0.1 units. $\text{goodnessOffit}$ metric is skewed to the left-hand side. The densities of $\text{IC50}$, $\text{EC50}$, $\text{GI50}$ and $\text{areaUnderCurve}$ are concentrated at the middle, which consists of a wide range of values.

Kernel density plots were plotted as well in order to understand the Kernel density estimation. It is a non-parametric method of estimating the probability density function (PDF) of a continuous random variable. Essentially, at every datum, a kernel function is created with the datum at its center – this ensures that the kernel is symmetric about the datum.
Results

The estimated regression coefficients that ran in the mentioned groups (pointed by cluster analysis) represent the rate of change in outcome units per unit change in the predictor variable.

Several runs of regression and clustering techniques were performed and it was possible to clearly identify a set of compounds with a clear advantage related to the dose-response curve. The results and proposed framework certainly will aid in developing effective collaboration and improved efficiencies for research groups within the life sciences industry.

We have chosen three kinds of regression models for detecting the best small molecules (drugs) based on the dose response data. The methods proposed include,

**Ordinary Least Squares** - Linear Regression fits a linear model with coefficients to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.

**Ridge Regression** - Ridge Regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares.

**Elastic Net Regression** - Elastic Net is a linear regression model trained with L1 and L2 prior as regularizer. This combination allows for learning a sparse model where few of the weights are non-zero like LASSO, while still maintaining the regularization properties of Ridge. We control the convex combination of L1 and L2 using the L1_ratio parameter.

For this phase of the project, we have implemented the Ordinary Least Squares model in detail.

As a preliminary step, a k-means clustering analysis was carried out in order to identify similar groups of small molecules’ data to be modeled. To determine the number of clusters, we used two approaches; WSS (within groups sum of squares) and Calinski criterion.

The plot below shows within groups sums of squares vs. the number of clusters extracted. The sharp decrease from 1 to 3 clusters (with little decrease after) suggests a 3-cluster solution.
To verify the number of clusters, Calinski criterion method proved that the optimal number of clusters must be 3. The criterion divides the data into partitions and compares them, as shown below.

Now the data from each of the three clusters must be incorporated into the modeling.

The sizes of the three clusters are 773,523,585 and their cluster means are shown below:

The cluster plot below shows the positioning of the three clusters within two components; component 1 and component 2. Clearly, the 2nd and 3rd clusters have a number of data points in common and the 1st cluster is isolated from the other two clusters.
By applying k-means to dose response data, we store the clustering result in a variable $k_c$. The plot below has IC50 (log10) on x-axis and maxDose (log10) on y-axis and shows the behavior of the three clusters with their individual cluster centers. We can notice that range of values for IC50 (log10) changes for every cluster.

As shown in the table below, there are some small molecules that belong to only one cluster, for example ‘(Z)-4-Hydroxytamoxifen’ is a small molecule that’s present only in Cluster 3, whereas ‘Gefitinib’ is another small molecule which is present in clusters 1 and 3 but not in Cluster 2.

<table>
<thead>
<tr>
<th>Small Molecule</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffeine</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dihydroxytamoxifen</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S-DFUR</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>S-FU</td>
<td>0</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>Afatinib</td>
<td>5</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>AG1478</td>
<td>9</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Baiacain</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Bortezomib</td>
<td>7</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>Bosutinib</td>
<td>25</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Carboplatin</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>CCC-11047</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CGC-11144</td>
<td>20</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Cisplatin</td>
<td>12</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>CPT-11</td>
<td>7</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Docetaxel</td>
<td>14</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>Doxorubicin</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Epirubicin</td>
<td>27</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Erlotinib</td>
<td>8</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Etoposide</td>
<td>21</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Everolimus</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Fasciavsin</td>
<td>25</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Gefitinib</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Geldenemycin</td>
<td>14</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Gemcitabine</td>
<td>19</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>GS1059615</td>
<td>42</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>GS1070916</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>GS143911A</td>
<td>14</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>GS1556456</td>
<td>2</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>GS1461364</td>
<td>6</td>
<td>22</td>
<td>2</td>
</tr>
</tbody>
</table>

**Ordinary Least Squares (OLS) Regression**

Linear Regression fits a linear model with coefficients to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.

When a linear pattern is evident from a scatter plot, the relationship between the two variables is often modeled with a straight line.
When modeling a bivariate relationship, $Y$ is called the response or dependent variable, and $x$ is called the Predictor or Independent variable.

The simple linear regression model is written: $Y_i = \beta_0 + \beta_1x_i + \varepsilon_i$

To discover the correlation between the various dose response metrics, we implemented a dendrogram using 'Pvclust', an R package for assessing the uncertainty in hierarchical clustering. It provides AU/BP p-values (Approximate Unbiased, Bootstrap probability). This helps in highlighting the significant clusters.

$E_{\text{inf}}$ and $E_{\text{max}}$ are the metrics that highly correlate in terms of distance. Similarly, EC50(log10) and GI50(log10) are highly correlated.
For selecting the attributes for regression, a scatterplot matrix was plotted considering the data from Cluster 1:

The relationship between IC50.log10 as an outcome and potential predictors of IC50.log10 among the variables "GI50.log10.", "EC50.log10.", "areaUnderCurve", "hillSlope", "E_max", "E_inf";

"goodnessOfFit", "maxDose.log10."

We see major correlations between E_max, E_inf and EC50, GI50. There’s a strong correlation between GI50 and IC50 as well.

There were various diagnostic plots (see below) that were used to check the heteroscedasticity, normality, influence and the residuals vs. fitted charts that give us an insight on the trend of residuals. The QQ-plot suggests that residual errors are normally distributed. The points lie pretty close to dashed line.

The lower right plot shows standardized residuals against leverage. Leverage is a measure of how much each data point influences the regression.
While running the OLS model in R, there are three aspects that are unjustifiably driving our data:

- **Outlier**: unusual observation
- **Leverage**: ability to change the slope of the regression line
- **Influence**: the combined impact of strong leverage and outlier status

The ‘rms’ package in R is used to run the OLS model using the formula here:

```
cluster1.ols <- ols(IC50.log10. ~ GI50.log10. + E_max + E_inf, data = cluster1)
```

We start by summarizing the distribution of the data with the `datadist()` function.

As we might have more than one data distribution object in the workspace, we need to then inform the `rms` package functions that we are specifically using the cluster1 data distribution by calling the `option()` function. We then run the `ols()` function to produce our regression model.

The result for the first chosen cluster looks like this below:
> cluster1.ols

Linear Regression Model

ols(formula = IC50.log10. ~ GI50.log10. + E_max + E_inf, data = cluster1)

<table>
<thead>
<tr>
<th>Model Likelihood</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio Test</td>
<td>Indexes</td>
</tr>
<tr>
<td>Obs 430</td>
<td>LR chi2 338.70</td>
</tr>
<tr>
<td>sigma 0.3123</td>
<td>d.f. 3</td>
</tr>
<tr>
<td>d.f. 426</td>
<td>Pr(&gt; chi2) 0.0000</td>
</tr>
<tr>
<td></td>
<td>g 0.378</td>
</tr>
</tbody>
</table>

Residuals

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0167</td>
<td>-0.19387</td>
<td>-0.01481</td>
<td>0.21603</td>
<td>1.86257</td>
</tr>
</tbody>
</table>

| Coef | S.E. | t     | Pr(>|t|) |
|------|------|-------|----------|
| Intercept -3.0491 | 0.1546 | -19.73 | <0.0001 |
| GI50.log10. 0.3796 | 0.0299 | 12.71 | <0.0001 |
| E_max 1.1548 | 0.1253 | 9.22 | <0.0001 |
| E_inf 0.1081 | 0.1347 | 0.80 | 0.4226 |
Discussion of Results

Who will utilize this research framework? Clinical research organizations and Pharma Industry Research decision makers within these organizations will understand the outputs of this work as a value added product.

What is unique about the proposed solution? The analytics will be readily available to the organizations so they can rapidly integrate various datasets in order to model their decisions. Time is money in this process and even small delays can translate to considerable money over the lifetime of the compound development.

The estimated regression coefficients represent the rate of change in outcome units per unit change in the predictor variable.

You could multiply the coefficients by a value for the predictor to get an estimate of the outcome for that value, e.g. IC50.log10 is higher by 0.3796 times each unit increase in GI50.log10.

The intercept (the regression constant or Y intercept) represents the estimated outcome when the predictors are equal to zero – the average outcome value assuming the predictors have no effect here, an IC50.log10 value of -3.0491.

The next steps for this project would be to:

1. Evaluate the initial results of the first technique proposed (Linear Regression) using data for cluster 2 and cluster 3 as well, similar to that of cluster 1.
2. Try to use PCA and SVM to improve clustering results.
3. Model using the other proposed techniques (Ridge Regression and Elastic Net regression)
4. Select the best small molecules results.
5. Compare the results with other related works and papers.
If you have specific questions about HMS LINCS data, software, and tool reuse, please email lincs-feedback [at] hms.harvard.edu

---

i Dose-response data were published in "Metrics other than potency reveal systematic variation in responses to cancer drugs" Fallahi-Sichani et al, 2013 and made publicly available by the Harvard Medical School LINCS Center.
Dose-response data were acquired from the Harvard Medical School LINCS Center, which is funded by NIH grants U54 HG006097 and U54 HL127365."

ii This reference was published in “Nature Chemical Biology” Fallahi-Sichani et al, (2013) Nature Chemical Biology and made publicly available by the Harvard Medical School LINCS Center.

---

Dell EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL EMC MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell EMC software described in this publication requires an applicable software license.

Dell, EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries.