**L1-regularized least squares regression to predict drug responses in breast cancer using molecular profiles**

**Arvind Sridhar\*, Jean-Paul Abbuehl$, Jonathan Bernard$, Krisztian Homicsko$ and Anguraj Sadanandam$,#**

**\***Embedded Systems Laboratory (ESL), Institute of Electrical Engineering, **$** Swiss Institute for Experimental Cancer Research (ISREC), Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

**#**Correspondence to Anguraj.sadanandam@epfl.ch

**Summary Sentence**: **L1-regularized least squares regression**

**Background/Introduction**

The subchallenge 1 aims to predict drug responses (GI50 values) in test samples using molecular features from training set of samples. We identified three major issues associated with this aim, and this motivated us to choose L1-regularized least squares regression. The issues are:

1. The different molecular profile data from distinct measurement platforms were diverse, and it was difficult to integrate them.
2. There were a large number of genes/parameters within each of the above six data sets (for eg. >35000 genes represented in the RNA).
3. There were a large number of missing data for many cell lines that were randomly spread in all these molecular and drug response data sets.

To overcome the above issue #1, we used statistical regression analysis that combines all the data sets irrespective of the molecular and platform differences and without preprocessing/normalizing individual data sets. Regarding issue #2, we selected relevant genes across samples from different data sets using various screening methodologies. Firstly, we included only those genes with standard deviation greater than 0.8 across samples in gene expression (both microarray and RNAseq) profiles. Later, we performed genomic identification of significant targets in cancer (GISTIC) [1] analysis to select those genes with significant DNA copy number changes in 44 breast cancer cell lines. Finally, we grouped together (metagenes) those genes that had correlation co-efficient greater than 0.75 across samples in methylation data set. We combined the selected genes from the above data sets and integrated the data with reverse phase protein lysate array (RPLA) data set for further analysis. We avoided exome data due to its complexity.

Finally, we used a random (-100) value that is not present in any of the datasets in place for missing values, wherever applicable. In addition, we chose this random value such that it did not compromise the drug prediction values to a greater level. This optimal approach deals with the missing data, while not compromising the goal. In cases where the GI50 values were missing in the training set of cell lines, we considered those as test cell lines.

**Methods**

The data sets were reduced as described above and combined in to a single set “combined-data set” (CDS, no preprocessing involved). Later, L1-regularized least squares regression was applied as described [2]. This form of training, in addition to providing “weights” for each gene/parameter in the CDS, also helped “sparsify” the parameters. The cost function reduced in this form of training is:

,

where *y* is the list of drug responses (GI50 values) provided for a particular drug in the training, *A* is the matrix of the CDS (where rows represent cell lines and columns represents all the genes/parameters), *x* is the final weights of each parameter to be determined, and *λ* is a penalizing coefficient. In other words, in addition to minimizing the distance between *y* and *Ax*, the algorithm also penalizes the 1-norm of *x*, thus, eliminating all, but the most significant parameters needed to predict *y*. If higher the value of the coefficient *λ*, a large number of *x* values (weights of genes that predict drug responses) will become insignificant, and *vice versa*. We used the implementation provided in [2], where relative tolerance can be provided to solve the L1-regularized least squares problem within a given residual. During each iterations of the regularization process, the least squares problem was solved using the preconditioned conjugate gradient (PCG) method.

**Conclusion/Discussion**

Initially, we used 18 training cell lines and their data to identify molecular markers of drug response and later, we predicted drug responses for 35 test cell lines. Finally, we combined the drug responses from the 53 training and test cell lines and ranked them. There was a compromise on the prediction results due to the assignment of random number for missing data as discussed in issue #3. A better solution instead of assigning a random number for missing data in issue #3 discussed above could improve the results.

We tested the performance of our algorithm using GI50 data from a panel of breast cancer cell lines treated with a drug, lapatinib (inhibits dual tyrosine kinase receptor) from our previous publication [3]. As expected, our analyses identified HER2 as a top scoring gene in the combined data set as well as individual data set. However, our algorithm showed that gene expression profiles were better in predicting the drug responses compared to that of the other molecular profile data. This is probably true as gene and protein expression are final determinants of drug responses.

Overall, in the post genomic era that generates high-throughput molecular and drug data, our algorithm performs drug response prediction analysis by integrating diverse data sources irrespective of different platforms being used.

**Authors Statement**

Aravind Sridhar conceived the idea of using this algorithm, performed the analyses, interpreted the data and co-wrote the description. Jean-Paul Abbuehl helped with data analysis and critical discussion. Jonathan Bernard and Krisztian Homicsko participated in critical discussion. Anguraj Sadanandam preprocessed the data for the analyses, interpreted the data, co-wrote the description and supervised the project.

**References**

[1] http://www.stanford.edu/~boyd/papers/l1\_ls.html

[2] Beroukhim, R.*, et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007-20012 (2007).

[3] Heiser, L.M.*, et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 2724-2729 (2011).