Transfer Learning on Gene Expression Data for Cancer Drug Response Prediction

Angelo Miskalis
Department of Bioengineering
University of Illinois at
Urbana-Champaign Illinois, USA
angelom3@illinois.edu

Siva Nalla
Department of Bioengineering
University of Illinois at
Urbana-Champaign Illinois, USA
snalla2@illinois.edu

Ananthan Nambiar
Department of Bioengineering
University of Illinois at
Urbana-Champaign Illinois, USA
nambiar4@illinois.edu

***Abstract*— Gene expression and dose response data has the potential to be utilized for predictive models in clinical samples in fields such as oncology. Gene expression data, however, tends to have high dimensionality, leading to the "curse of dimensionality" when developing models to predict drug response in diseases such as cancer. Higher dimensionality in expression data requires more *in vitro* experimentation, leading to an increase cost in cell lines and reagents. To mitigate this, we produced a lower dimensional embedding for gene expression data. We utilized the cancer genome atlas (TCGA), which provides a robust amount of gene expression data from 10,000 cell lines and 33 cancer types. We investigated three different methods to produce the embeddings. The first used a convolutional neural network to classify gene expression data from the 10,000 cell lines to their cancer types. Then, one of the intermediate layers from these networks were used to produce feature vectors that correspond to the gene expression data. The second is a convolutional autoencoder. The third method was principal component analysis (PCA), the benchmark technique for dimensionality reduction in gene expression data. These low dimensional feature vectors were then utilized to predict the drug response for cell lines via a logistic regression model. We compared the accuracy our logistic regression model to that of a state-of-the-art random forest based method, and with the convolutional neural network, obtained an accuracy within ten percent of the former model. The other methods possessed decreasing accuracy. These results indicate that cancer type prediction tasks encode generalizable features for cell lines, which can be leveraged to create less computationally straining and cost-effective models to predict drug response. Related code is made publicly available at : https://github.com/annambiar/GeneExp**

Keywords— Transfer Learning, Deep Learning, Logistic Regression, Gene Expression Analysis, Dimensional Reduction, Principal Component Analysis, Autoencoding, Convolutional Neural Networks, miRNA, Clinical Oncology

I. Introduction

Gene expression, specifically RNA expression, is used as a robust tool in clinical medicine to measure therapeutic efficacy. RNA's consist of multiple classes such as siRNA, mRNA, tRNA, and miRNA. The latter were discovered in the late 1980's as a robust biomarker that circulates freely in the intra and extracellular environments. In the last two decades, it has been shown that miRNAs play a major role in critical feedback loops. miRNA, which is usually 22 base-pairs in length, regulates the translational ability of mRNA [1]. This is achieved via binding to mRNA sequences, which prevent translation; miRNA binding may also recruit proteolytic enzymes to digest the mRNA. This understanding that miRNA's has a role in regulating the mRNA's translation of proteins is crucial because it allows us to monitor the gene (miRNA and mRNA) expressions of many types of healthy and diseased cells. Gene expression data from diseased cells is certainly more valuable than the protein concentrations because it allows us to understand the initial genetic source or cause of protein production and inhibition regulation. Furthermore, many studies over the last couple decades have correlated increased levels of miR's to downregulation of mRNAs and protein productions in human cancer cells. It has also been reported that increased levels of cancer associated miRNA's in patient blood samples can lead to limited therapeutic responses and detrimental fates to the survival rates of patients [1].

Commonly, miRNAs and nucleic acids appear in low quantities in extracellular environments, causing sensitive experimental detection difficult which yield large statistical variations in gene expression data. Gold-standard techniques such as Real-Time quantitative polymerase chain reaction (qPCR) generate lower limits of detection (LOD) in the picomolar-femtomolar range [2]. Techniques with similar LODs include Next Generation Sequencing (RNAseq). Along with miRNAs, mRNAs can be read with the aforementioned techniques. These sequencing techniques do possess drawbacks. Gathering and generating data is lengthy and complex. This requires lengthy and precise extraction of RNA. Then, the RNA is converted to cDNA using reverse transcriptase (**Figure 1**). The converted cDNA is then sheared into fragments where similar length fragments are ligated to known sequence adapters for high accuracy directional sequencing (or map reads)[2]. This can yield biases between various experimental protocols for gene expression. Kits to perform these experiments are also expensive. For example RNAseq costs approximately $175.00 to collect a large set of gene expression data from

one sample [3]. These complications can make the formation of lengthy gene expression profiles for multiple cell lines difficult.

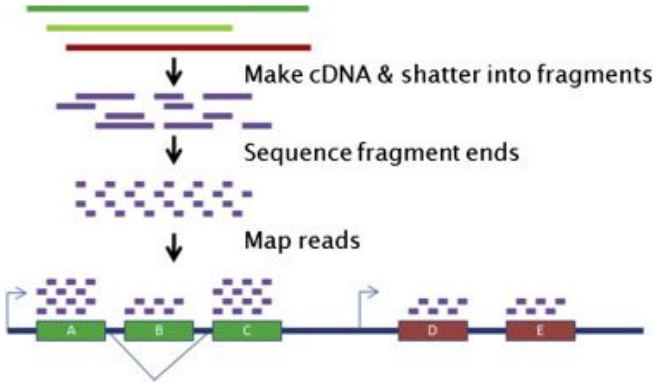

***Figure 1**: Schematic illustrating the preparation of samples RNA for an RNAseq experiment. [2]*

There is, however, merit to genomic analysis in data modeling. Data from expression profiles has been utilized to formulate correlations between gene expression and cellular functions such as viability and proliferation [3]. Thus, there is an interest to utilize these correlations as prediction models. An area of recent interest is the prediction of drug efficacy based on gene expression data[4]. In clinical settings, it is routine to measure drug efficacy through dose response curves (**Figure 2**). A dose response curve is often fitted to a standard Hill model, and either the area under the curve (AUC) or half-maximal dose concentration (EC50) is used to describe drug efficacy [5]. In prediction models, gene expression is utilized to estimate drug efficacy in either a binary or more fine-grained manner.

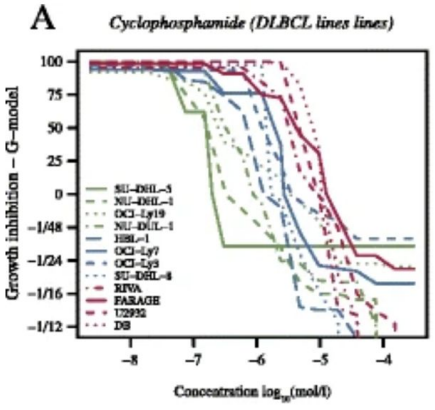

***Figure 2:** Representative dose response curve [4]*

Obtaining large sets of gene expression data and using it to predict drug response is useful but challenging. Aside from the aforementioned experimental complications, gene expression datasets possess high dimensionality, leading to the "curse of dimensionality", where the amount of data required increases exponentially with the number of dimensions. This is also computationally straining, requiring the use of supercomputing systems or hours of computation. Therefore, it is of interest to produce a lower dimensional embedding which possess prediction accuracies similar to higher dimension systems. This would reduce computational cost, and could also help better identify genes of interest rather performing an expensive screen of all genes.

In this study, we explored a large set of gene expression data from various types of cancer cell lines to design transfer learning and crosscorrelation computational models to predict cancer drug response without high computational strain or time and cost. We utilized The Cancer Genome Atlas (TCGA) to analyze gene expression data from over 33 cancer types and 10,000 cell lines, and developed a computational tool to bring this large set of data into a more manageable lower-dimensional gene expression data set. Utilizing the lower dimensional gene expression dataset we applied logistic regression modeling on EC50 drug-dose response data gathered for 5 cancer types from the Maslov group at UIUC to predict cancer drug response [5]. Results of these models were compared with that of principal component analysis (PCA), a common dimensionality reduction technique in gene analysis. We theorize that gene expression to drug response correlations from this study will provide insights into the role of mRNA and miRNA expression levels to drug response. Future studies providing gene expression response data to drug therapy at various time-scales can be utilized to develop better clinical therapeutic regimens. When these time scale data are combined with large random biospecimen gene expression data, we can get a better understanding of cellular mRNA and miRNA functional similarities and differences amongst a variety of cellular sample data, further allowing us to develop better drugs for cancer cell fate.

II. Data

This study utilized a large dataset with over 10,000 cell lines and 33 cancer types from The Pan Cancer Atlas [6]. For each cell line, the dataset includes the RNAseq counts of over 20000 genes and the type of cancer for the cell line.

Drug effects on cancer cell lines are explored via the CTRP database which has been processed to reduce noise produced by high throughput experiments as a part of an upcoming paper by Argonne National Lab. The data used in this study consisted of gene expression profiles for each cell line with their respective AUC values. These were binary AUC values, where 1 indicated cells that were affected by the drug. The cutoff for this binary classifier was the AUC. an area lower than 0.5 classified the cell as 1; all others were set to 0. This study focused on the drug Paclitaxel, for which drug response data for about 700 cell lines were available. The cell lines possessed expression profiles for 16,383 genes.

Lower dimensional features were developed using the Pan Cancer Atlas dataset and logistic regression was applied to these feature vectors and paired with CTRP drug response data to predict drug effectiveness.

Since the range of values for RNAseq counts easily goes over two orders of magnitude, the log transform $\ln(x+1)$ is used. Furthermore, only genes that are present in both the

Pan Cancer Atlas and CTRP database are kept to allow for transfer learning to occur. Finally, the gene expression vectors are reshaped into expression matrices because it has been previously shown that reshaping has been shown to be an effective representation of gene expression data for deep learning (**Figure 3**) [7].

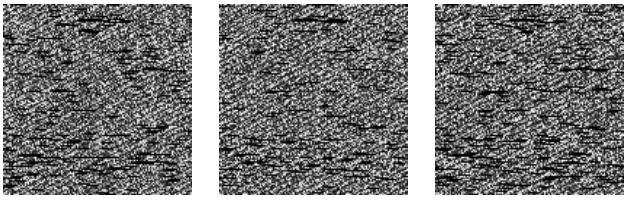

Figure 3: *A sample of preprocessed gene expression grids*

III. Model and Methods

The models proposed in this paper consists of two parts as shown below (**Figure 4**). The first part of the model involves a deep neural network and the second part is made up of a simple logistic regression classifier. These two parts perform different learning tasks. First, the deep learning module of the model utilized the large dataset made available from the Pan Cancer Atlas to perform either a supervised or unsupervised auxiliary task, $\vec{y'} = f(\vec{x})$ where $\vec{x} \in \mathbb{R}^{16383}$ is the gene expression vector for a particular cancer cell line, f is the deep learning algorithm and $\vec{y}$ is the output for the deep learning task. As mentioned in the previous section the vector $\vec{x}$ was reshaped into a two dimensional grid to enable the utilization of a vast variety of deep learning algorithms originally developed for image analysis. For the second step, the deep learning model was used to produce a lower dimensional embedding $\vec{x'}$ for the gene expression vectors of the cell lines from the CTRP database. which was then used as input for the logistic regression classifier $\vec{y} = g(\vec{x'})$ where $\vec{y} \in \{0,1\}^2$ encodes whether or not the drug was effective.

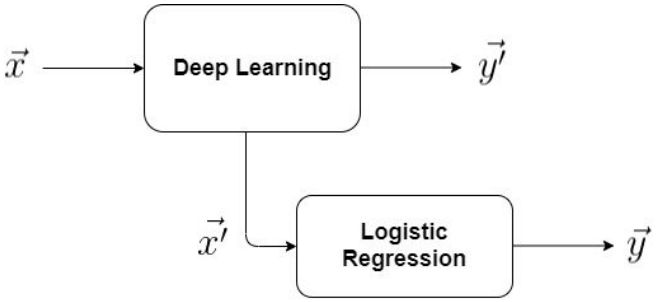

Figure 4: *The workflow outline for transfer learning models*

For this workflow, two different deep learning algorithms were tested. The first method, ConvCType, is a convolutional neural network classifier that is trained with the auxiliary task of classifying gene expression data of cell lines into one of the 33 different cancer types available in the Pan Cancer Atlas. The architecture for the neural network, which was inspired by Lyu and Haque [6] is as shown below (**Figure 5**). The input to ConvCType is a two dimensional representation of the gene expression vector. The first two layers of the neural network are convolutional layers with kernels of size 5 followed by another convolutional layer with kernel size 3. This is followed by a fully connected layer with a ReLU function. This hidden layer is used to produce the embedding $\vec{x'}$ that were used in the following step by logistic regression (**Figure 4**). The final layer of ConvCType is a fully connected layer with the softmax. The final layer performed the auxiliary task by predicting the probability that the input is of a particular type, represented by the vector $\vec{y'}$. The argmax of these probabilities could then be used to assign a cancer type to the input. This neural network is optimized by minimizing the cross entropy loss of the classifier.

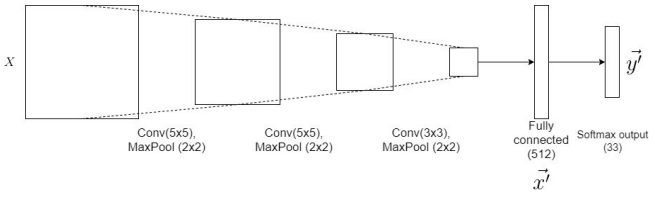

Figure 5: *ConvCType, a convolutional neural network based cancer type classifier*

Note that ConvCType extracts features that are important for a specific biological task (cancer type classification) that might not be directly related to the task of drug response prediction. Therefore, to compare it to a method that extracts more general features, the autoencoder network shown below (**Figure 6**) was also used to produce low dimensional embeddings [8]. The autoencoder network has first two layers are convolutional layers that are similar to ConvCType. However, this is followed by two upsampling layers using transposed convolutional kernels. Essentially, the autoencoder network compresses the data and then tries to retrieve the original data point. The compressed data X' was then used by the logistic regression algorithm to predict drug response. To train the weights of the autoencoder, the mean squared loss was used along with a stochastic gradient descent optimizer.

Both neural networks were implemented in Python using the PyTorch library and were trained on a machine with 4 NVIDIA 1080 Ti GPUs for 10 epochs.

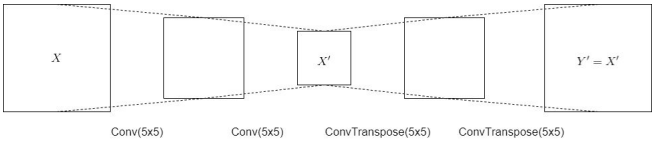

Figure 6: *A convolutional autoencoder neural network*

The third technique, principal component analysis (PCA), utilized gene expression data obtained from the modified CTRP database. Each of the 773 cells lines had their gene expression profiles reduced into their principal components. This was done through MATLAB to perform the task $\vec{x'} = f(\vec{x})$, where $\vec{x}$ represents the gene expression profiles, and f is MATLAB's pca function. This function will take an *n x p* input, where *n* is the variable and *p* is the condition, and generate a *p x p* output. This generates *p* principal components. Therefore, this function took in the

cell and expression data, and output the first 773 principal components. For a comparison analysis, the first 500 principal components were used.

MATLAB was utilized to make a logistic regression model for the three classification methods. This model was used to predict if a cell line was either affected by Paclitaxel. All rows were shuffled before predictions were calculated. Data was split into training and test groups which possesses 70% and 30% of the data, respectively. The hypothesis formula to fit this prediction model followed a sigmoidal function (**Equation 1**). A cross-entropy cost function and gradient vector implemented for optimization (**Equations 2, 3**).

$$h_\theta = \frac{1}{1 + e^{-\theta x}}$$

Equation 1: *Sigmoidal hypothesis function for logistic regression model. Theta = estimated parameter, x = gene expression data*

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y(i)\log\left(h_\theta(x(i))\right) + (1 - y(i))\log(1 - h_\theta(x(i))]$$

Equation 2: *Cost function. n = cell line number, y = actual AUC value, h = hypothesis function with parameter theta, x = gene expression data*

Optimization of the logistic regression fit was solved via minimization of the cost for the parameters of the sigmoidal hypothesis function. This was performed in MATLAB via the fminunc command, which inputs the gradient vector and cost function to find a theta which reaches a cost closest to zero. The function was run 400 times, in which the function found a local minimum prior to finishing its iterations.

$$grad(i) = 1/n\sum[(h_\theta(x_i) - y_i)x_{ij}]$$

Equation 3: *Gradient function. n = cell line number, x = gene expression data, h = hypothesis function with parameter theta, y = actual AUC value*

Following optimization, AUC predictions were calculated for the training and test groups of the three methods through MATLAB's predict function. This prediction was performed twenty-five times to generate a mean dataset. The mean accuracy for all methods were reported in percentage of correct decisions. Mean true positives, true negatives, false positives, and false negatives were also reported as overall percentages.

IV. RESULTS

The aforementioned classification methods were compared to that of a state of the art random forest based classifier from an upcoming paper by Nambiar, Dubinkina and Maslov [9]. The Interpret Forest method possessed the highest accuracy, followed by the ConvCType, Autoencoder, and PCA techniques (**Table 1, Figure 7**).

Table 1: *Accuracies for drug response achieved by various classifiers*

ConvCType + logistic	0.63
Autoencoder + logistic	0.55
PCA + logistic	0.47
Interpret Forest	0.73

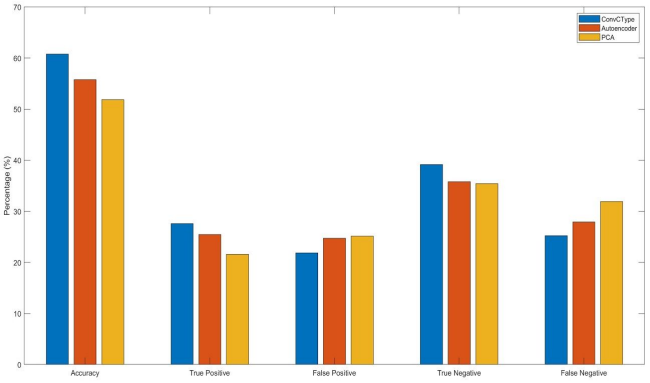

Figure 7: *Mean accuracies, true positive, false positive, true negative, and false negative values for each classification method.*

It was seen that although, ConvCType+logistic and Autoencoder+logistic do not outperform Interpret Forest, the former was within ten percent of it and the latter fifteen. Importantly, they were better than the benchmark.

Along with mean accuracies, comparisons were made between the true/false positives and true/false negatives between the groups analyzed via logistic regression (**Figure 7, Table 2**). From these results, it was demonstrated that ConvCType based regression produced the highest true positive and negatives, followed by groups which had decreasing accuracy. False positive and negative groups showed a similar trend. There was also consistently in trends between negative and positive results.

Table 2: *True positive/negative and false positive/negative values*

Method	True Positive (Mean %)	True Negative (Mean %)	False Positive (Mean %)	False Negative (Mean %)
ConvCType	27.61	39.15	21.85	25.23
Autoencoder	25.48	35.8	24.75	27.92
PCA	21.6	35.44	25.15	31.91

ConvCType+logistic and Autoencoder+logistic perform significantly better than PCA+logistic. This indicates that there is merit to using a nonlinear dimensionality reduction technique. This is in line with expectations as biological processes tend to be nonlinear.

Moreover, ConvCType+logistic outperforms Autoencoder+ logistic in most parts of the confusion matrix. This demonstrates that using a biologically relevant auxiliary task to obtain embeddings provides an advantage to the

classifier. Furthermore, it also shows that the features that were learnt for cancer type classification are relevant for drug response prediction.

V. Discussion

With recent advances in transcriptomics, there arises a need for methods to efficiently store and utilize large datasets of gene expression data. This is due to the high dimensionality of gene expression data. This paper studied methods for embedding gene expression data in a lower dimensional space with a focus on nonlinear dimensionality reduction using deep learning. The dimensionality reduction was then tested by performing the task of drug response prediction.

While the methods for dimensionality reduction, ConvCType+logistic and Autoencoder+logistic, did not outperform the state of the art drug response prediction algorithm, with further finetuning of the deep neural network architecture, this gap could be decreased. Furthermore, it is important to note that ConvCType+logistic and Autoencoder+logistic use a simple logistic regression to make the classification decision. This was done to highlight the power of the embedding itself when making the discrimination. However, using a more complex classifier with the embedding could also be beneficial.

The true positive/negative and false positive/negative analysis indicated that the three methods produced trends that correlated with their accuracies. Additionally, to further improve the method analysis, a receiver operator characteristic (ROC) curve could be implemented. This would allow for better classification of the sensitivity of each method.

Finally, the observation that the biologically motivated auxiliary task used by ConvCType does a better job at providing the features relevant for drug response prediction than Autoencoder is interesting and warrants further computational study into the relationship between cancer type and drug response.

VI. Conclusions

In this paper, we studied methods for dimensionality reduction of gene expression data and tested these methods by predicting cancer drug response. It was found that nonlinear methods for dimensionality reduction significantly outperformed linear dimensionality reduction. Furthermore, it was found that dimensionality reduction using transfer learning with a biologically motivated auxiliary task performed better than a more general auxiliary task.

VII. Contributions

The authors of this paper contributed equally to the regression model concept development, true/false positive and true/false negative analysis, and paper writing. Siva worked on the background significance of gene expression, clinical drug-dose response, analytical assays, and data sections. Ananthan utilized the Convolution Neural Networks (CNN) and developed an autoencoder algorithm for analyzing the TCGA data using Python. He was also responsible for doing data compilation and preprocessing. Angelo focused on the analysis of the drug response data and developed Matlab code using a logistic regression modeling; he also wrote the code which performed Principal Component Analysis (PCA). All logistic regression-based plots were implemented in his code. All members were equally involved in the results analysis and discussion of this work.

VIII. References

[1] N. Kosaka, H. Iguchi, T. Ochiya. "Circulating miRNA in body fluid: a new potential biomarker for potential cancer diagnosis and prognosis." *Cancer Science.* 2010.

[2] S. Seesi, A. Kueck, *et al.,* "Genomics guided immunotherapy of Human epithelial ovarian cancer." *Translational Cardiometabolic Genomic Medicine*. 2016.

[3] Hughes, T. "Validation in Genomic Scale Research." *Journal of Biology.* 2009.

[4] Steffan, G., *et. al.*, "Predicting Response to Multidrug Regimens in Cancer PAtients Using Cell Line Experiments and Regularized Regression Models.", *BMC Cancer*, 2015.

[5] G. Veroli, C. Fornari, Ian Goldlust *et al.* "An automated fitting procedure and software for dose-response curves with multiphasic features." *Scientific Reports.* 2015.

[6] J. Weinstein, *et al.,* "The Cancer Genome Atlas Pan-Cancer analysis project", *Nature Genetics*, 2013.

[7] B. Lyu and A. Haque, "Deep Learning Based Tumor Type Classification Using Gene Expression Data", *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, 2018.

[8] M. Chen, X. Shi, Y. Zhang, D. Wu and M. Guizani, "Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network", *IEEE Transactions on Big Data*, 2017.

[9] A. Nambiar, V. Dubinkina, S. Maslov. "Interpretable Random Forests for Cancer Drug Response Prediction." *In preparation*, 2019.