Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration

Dibyendu Bikash Seal[a], Vivek Das[b], Saptarsi Goswami[c], Rajat K. De[d,*]

[a] A. K. Choudhury School of Information Technology, University of Calcutta, JD-2, Sector III, Salt Lake City, Kolkata 700106, India
[b] Novo Nordisk Research Center Seattle, Inc., 530 Fairview Ave N # 5000, Seattle, WA 98109, United States
[c] Bangabasi Morning College, 35 Rajkumar Chakraborty Sarani, Scott Ln, Kolkata 700009, India
[d] Machine Intelligence Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India

ARTICLE INFO

Keywords:
Gene expression
Denoising auto-encoder
Multilayer perceptron
Regression
Multi-omics integration
DNA methylation
Copy number variation

ABSTRACT

Gene expression analysis plays a significant role for providing molecular insights in cancer. Various genetic and epigenetic factors (being dealt under multi-omics) affect gene expression giving rise to cancer phenotypes. A recent growth in understanding of multi-omics seems to provide a resource for integration in interdisciplinary biology since they altogether can draw the comprehensive picture of an organism's developmental and disease biology in cancers. Such large scale multi-omics data can be obtained from public consortium like The Cancer Genome Atlas (TCGA) and several other platforms. Integrating these multi-omics data from varied platforms is still challenging due to high noise and sensitivity of the platforms used. Currently, a robust integrative predictive model to estimate gene expression from these genetic and epigenetic data is lacking. In this study, we have developed a deep learning-based predictive model using Deep Denoising Auto-encoder (DDAE) and Multi-layer Perceptron (MLP) that can quantitatively capture how genetic and epigenetic alterations correlate with directionality of gene expression for liver hepatocellular carcinoma (LIHC). The DDAE used in the study has been trained to extract significant features from the input omics data to estimate the gene expression. These features have then been used for back-propagation learning by the multilayer perceptron for the task of regression and classification. We have benchmarked the proposed model against state-of-the-art regression models. Finally, the deep learning-based integration model has been evaluated for its disease classification capability, where an accuracy of 95.1% has been obtained.

1. Introduction

Gene expression (GE) profiling allows capturing the genetic and epigenetic alterations that an organism undergoes under various biological conditions during its lifetime. Changes in it lead to different phenotypic traits. GE is affected by several factors including genetic factors like copy number variation or copy number alteration (CNV/CNA), DNA mutations and epigenetic factors like DNA methylation (DNAm) or histone modifications. Until recently, genetic and epigenetic changes have been considered separate events in cancer. However, recent studies have shown that they interweave together during tumor growth and progression [1]. Growing evidences have also shown epigenetic changes leading to genetic mutations and vice versa. DNAm occurring around promoters are believed to be linked with CNV and GE changes to understand the complex mechanisms behind cancer development and progression [2,3]. Thus, both genetic and epigenetic events may lead to abnormal gene expression. Methylation of CpG islands in cancer cells often leads to silencing (hyper-methylation) or activation (hypo-methylation) of gene expression due to blockage in the promoter regions (regions facilitating gene transcription), thereby restricting transcription factors to bind to the regions [4]. CNV, on the other hand, may also lead to the activation of certain oncogenes or silencing of tumor suppressor genes in cancer whereby an entire protein coding sequence in a given DNA can be either deleted or duplicated, thus affecting gene expression. Such changes may ultimately lead to more or less production of downstream protein leading to tumor phenotype [5–7].

Hepatocellular carcinoma (HCC) is one among the leading causes of cancer in the world. Patients with chronic liver diseases like fatty liver and cirrhosis are primarily affected by HCC. Recent research has shown that altered DNAm and CNV are two of those early events that take place during the pre-neoplastic phase of HCC, and are important in

* Corresponding author.
E-mail address: rajat@isical.ac.in (R.K. De).

https://doi.org/10.1016/j.ygeno.2020.03.021
Received 28 November 2019; Received in revised form 17 March 2020

classifying regenerative modules into distinct classes [8,9]. Hence, the role of genetic and epigenetic events in HCC needs attention. Although being linked with high mortality rates, HCC is relatively understudied and the problem of identification of biomarkers for the prognosis of liver hepatocellular carcinoma (LIHC) still needs to be addressed.

Recent years have shown an explosion of data due to Next Generation Sequencing (NGS) technology. This has led to the production of large paient specific data cohorts of various multi-omics types, including genomic, transcriptomic, epigenomic, proteomic, metabolomic, fluxomic, lipidomic, ionomic along with the corresponding clinical information, much of which have been accumulated in the international consortiums like TCGA (now moved to GDC) and ICGC. Encyclopedia of DNA Elements (ENCODE) project [10] is another such effort that annotates and maps functional elements across the genome. With the growth of multi-omics data, newer unexplored avenues have opened up that was lacking earlier by exploiting individual omic layer data to understand tumor biology. An integrative analysis of such multi-omics data can play a significant role in tumor diagnosis since they are reticular in nature, and can exhaustively portray the molecular variations that an organism undergoes during various stages of carcinogenesis.

Several efforts to integrate omics data from different modality in HCC patients have identified subtype characteristics associated with poorer prognosis and potential therapeutic targets [11]. Ecomics, a well-annotated and normalized multi-omics conspectus for *E. coli*, and Multi-Omics Model and Analytics (MOMA) platform built to learn from Ecomics, allow integration of four omics layers to design genome-wide models [12]. miRNA and mRNA expression data for pancreatic cancer have been integrated using machine learning models to evaluate single and multi-biomarkers for their diagnostic performance [13]. Penalized regression methods have been used to integrate omics data from three platforms to identify genes associated with both SNPs and CpGs [14]. A novel Multiple Kernel Learning (MKL) method has been used in [15] for prediction of prognosis of Gliobastoma Multiforme (GBM) by integrating histopathological images with multi-omics data. MKL-based methods have also been used to integrate genomic data with pathological images to predict clinical outcomes for breast cancer patients [16].

Integrative clustering methods of multi-omics datasets have been used profoundly to classify cancers based on molecular alterations [17]. Integration of omics platforms is not trivial as it involves preparing data from heterogeneous sources derived from various equipment and experimental settings. The present state-of-the-art pre-processing techniques including normalization, imputation and quality evaluation have been reviewed in [18]. DIABLO [19], on the other hand, is a multivariate dimension reduction method for multi-omics data, which is capable of identifying novel molecular signatures. The review articles in [20,21] discuss various challenges in integrating such multi-modal data with robust statistical precision.

An inherent feature of multi-omics data layers is that they contribute towards establishing a complex, non-linear structure. On a similar note, a deep neural network (DNN) architecture also resembles such layered non-linearity, where output values are computed sequentially, layer after layer. This allows complex useful features to be learnt in an unsupervised manner by combining multiple simpler features following a layered abstraction. Deep learning (DL) based-integration methods on TCGA data have been used earlier to identify survival subgroups linked with poor prognosis and major signaling pathways [22,23], finding influence of driver genes on phenotypes [24]. A multimodal DL model, employed for prediction of prognosis of breast cancer by integrating genomic and clinical data, has shown better performance than usual methods not integrating omics data [25]. A random forest model integrating CNV, DNAm, microRNA and transcription factors (TFs) has been used to predict gene expression for Liver HCC patients [26]. A micro-array based DL model D-GEX [27] has been shown to outperform linear regression models in predicting GE of landmark genes. TCGA DNAm data and specific histone ChIP-seq data have been integrated to predict GE in lung cancer using ReliefF for feature selection and random forest for classification [28]. Another DL model using deep auto-encoders and multi-layer perceptrons (MLP) has been used in [29] to predict GE using genetic variants on yeast dataset. This MLP-SAE model has been shown to outperform commonly used standard models like Lasso and Random Forest. Nevertheless, research is limited as no robust DL model exists that takes into account the effect of both genetic and epigenetic perturbations while predicting non-linear GE function [30], giving better characterization of the disease. Multiple other efforts have already shown the impact of DNAm on GE [31]. A similar investigation has been carried out to estimate GE based on DNAm profiles for breast cancer using a L1-regularized regression model [32].

Thus, current exponential growth of varied NGS technologies and compelling evidence as cited above have already provided enough clue towards genomic and epigenomic factors influencing or guiding gene expression in a tissue-specific manner giving rise to diverse phenotypic traits. Single omics analysis using either CNV or DNAm data to predict gene expression has already been carried out in [33–35]. However, with growing evidence, we are arriving at a point of realization that gene regulation in cancer phenotypes is not entirely driven or influenced by a single factor. Taking a cue from all these published evidence, here we have tried to estimate gene expression in LIHC as a function of DNAm and CNV at protein-coding regions.

In order to perform the task of estimating gene expression, we have developed a DL regression model based on multi-omics integration. We have used deep denoising auto-encoder (DDAE) for feature extraction and multi-layer perceptron (MLP) for regression. Auto-encoders have found significant usage in feature extraction and dimension reduction in recent years [36–38]. In the present study, we have intended to extract features enough to explore the relationship between genomic (CNV) and epigenomic (DNAm) information in regulating gene expression at higher dimensions. The proposed predictive model filters signals from noise contributed via both these genomic and epigenomic platforms, understands the non-linear relationships among the input features, and finally captures the influence of these relationships to extract information encoded in mRNA expression for paired sets of patient samples. It may be mentioned here that current LIHC TCGA studies involving multi-omics integration have often been limited by sample size as outlined in [11]. In this work, we have used 404 paired samples. The DDAE-MLP model has shown comparable performance against state-of-the-art regression methods.

2. Materials and methods

This section explains the methods used for data acquisition, pre-processing and the deep learning-based methodology used in this work to estimate gene expression from DNAm and CNV data.

2.1. Data acquisition

TCGA multi-omics data for Liver Hepatocellular Carcinoma (LIHC) from TCGA portal (now moved to Genomic Data Commons https://gdc.cancer.gov/) have been used in this work. The R package TCGA-assembler (v2.0.5) [39] has been used to obtain DNAm, CNV and RNA-seq data for LIHC. The number of samples for each omics type and their corresponding assays is listed in Table 1:

2.2. Pre-processing

For the DNAm data, we have calculated the average methylation values by mapping CpG islands within 1500 bps from the transcription start site (TSS) (both DNAse hypersensitive and hyposensitive). For all three omics data, the common samples (patient Identifiers) have been identified. Each omics data has been reduced to the common samples

Table 1
Number of samples and their corresponding assays for each omics type.

Omics type	#Samples	Assay	Access level	Platform
DNA methylation	429	Infinium HumanMethylation450 BeadChip	3	methylation_450
CNV	424	Affymetrix SNP Array 6.0	3	cna_cnv.hg19
RNA-Seq	761	Illumina HiSeq	3	gene.normalized_RNA-seq

Table 2
Dimension for each omics type after pre-processing.

Omics type	Number of samples	Number of features (genes)
DNA methylation	404	18,996
CNV	404	23,604
RNA-seq	404	15,397

only. As a pre-processing step, for all three omics data, we have performed the following, as suggested in literature [40]. First, the genes which have more than 20% missing values across all samples (patients) have been removed. Secondly, the samples which have more than 20% missing values across all features (genes) have been removed. These steps resulted in 404 common samples. The package sklearn.preprocessing.Imputer (Sckit-learn) [41] has then been used to impute the remaining missing values using the mean across all features (genes). Finally, each omics data has been normalized and scaled in [0,1] range using sklearn.preprocessing.MinMaxScalar (Sckit-learn) package [41]. Additionally, in order to remove features with zero or relatively low expression values, we have calculated the rowsum for each feature and removed the first quantile for RNA-seq data. The dimensions of the finally pre-processed omics data are listed below in Table 2.

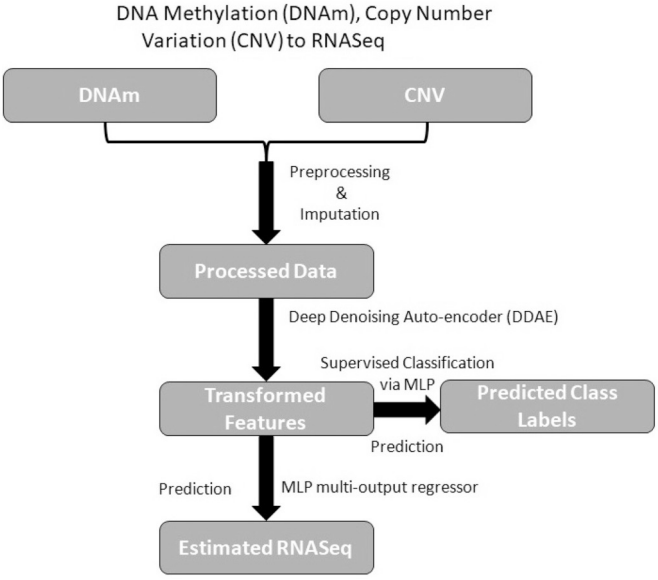


Fig. 1. Overall workflow of the deep learning-based regression model.

2.3. Deep learning-based methodology

In this article, we have developed a deep learning regression model for multi-omics integration that uses a DDAE network for feature extraction and dimensionality reduction, and a MLP network for regression. First, we have built a deep learning-based regression model for estimating gene expression from DNA methylation and copy number variation profiles. Finally, we have performed classification of tumor and normal samples based on the features extracted using the deep learning framework to validate the proposed model. The overall workflow of the deep learning-based regression model showing all experiments carried out in this work has been depicted in Fig. 1.

Auto-encoders (AEs) have been found to have profound usage in applications involving feature extraction and dimensionality reduction as mentioned in Section 1. Both Principal Component Analysis (PCA) and Auto-encoders are unsupervised techniques which aim at minimizing the reconstruction loss. Auto-encoders can theoretically perform better than PCA as they do not have any restriction on linearity. In a recent work [42], even for special cases like highly unbalanced datasets, auto-encoders have performed better than PCA. Similarly, another investigation highlights that auto-encoder yields better result compared to linear PCA on specialized tasks like anomaly detection without needing the computational overhead of Kernel PCA [43]. In [44], the authors have revealed the capability of AEs to find repetitive structures as compared to other methods including PCA and Linear Discriminant Analysis (LDA). The facts discussed above have motivated us to use AE based feature representation and reduction over PCA for the current work.

2.3.1. DDAE-MLP model for estimating gene expression from DNA methylation and copy number variation data

In this work, we have attempted to establish the fact that GE alterations can be captured by integrating a genetic factor like CNV and epigenetic factor like DNAm. Here, we have modelled the problem of estimating gene expression in two ways. In one approach, DNAm data and CNV data have been scaled and stacked together to form the integrated input for the DDAE network. The DDAE network has been used to extract significant features from the integrated data. DDAE model consists of an input layer, an output layer and two auto-encoders in between. The DDAE model has been trained using stochastic gradient descent algorithm with MSE as the loss function.

After training, the reduced features have been extracted from the bottleneck layer of the DDAE. These features have served as the input for the MLP-based regression model. The output of the regression layer has the same number of neurons as the number of target variables (genes) in the RNA-seq data. MLP-regression model consists of three hidden layers with ReLU activation function. The output of the regression model uses linear activations. MLP has been trained with an 'adam' optimizer using MSE as the loss function. The architecture of the DDAE network and the MLP-based regressor network have been shown in Figs. S1 and S2 (in Supplementary Material) respectively. The block diagram of the DDAE-MLP model used in this work for estimating gene expression has been depicted in Fig. S3 (in Supplementary Material).

The second approach is to build a DDAE for each source DNAm and CNV first, and then concatenate the extracted features to get the input for the regression layer. However, this method has not produced any better result as compared to the method mentioned above. 'DDAE-MLP individual' represents the results obtained from the second modelling approach in Fig. S4 (in Supplementary Material).

2.3.1.1. Feature extraction using deep denoising auto-encoder. The features from the input data (DNAm and CNV stacked together) have been extracted using a DDAE network. An auto-encoder [45] is a non-recurrent, feed-forward neural network that consists of an input layer, one or more hidden layer(s) and an output layer. The number of neurons in the output layer is the same as that in the input layer since the output is an approximate reconstruction of the original input. An auto-encoder network typically employs two functions: an encoder

function $\mathbf{y} = \mathbf{u}(\mathbf{x})$ and a decoder function $\mathbf{z} = \mathbf{v}(\mathbf{y})$. The output $\mathbf{x}'$ of an auto-encoder is simply a reconstruction of the original input such that $\mathbf{x}' = \mathbf{v}(\mathbf{u}(\mathbf{x}))$. The number of neurons in the hidden layer is kept smaller than the number of neurons in the input layer in order to improve generalization capability, and hence the dimensionality of the input data is reduced.

The input to the DDAE network is the combined data having 404 samples and 18996 (from DNAm) and 23604 (from CNV) making a total of 42600 (say n) features. The auto-encoder thus transforms an input $\mathbf{x} \varepsilon \Re^n$, through a series of hidden layers to $\mathbf{x}' \varepsilon \Re^n$ such that $\mathbf{x}' \approx \mathbf{x}$. We have used ReLU (Rectified Linear Units) activation function for each hidden/output node following McCulloch-Pitts model of neuron. An auto-encoder is trained to minimize the reconstruction error $\|\mathbf{x} - \mathbf{x}'\|^2$. We have used a denoising auto-encoder [46] where the input data is first corrupted by adding noise elements to it and the corrupted input $\tilde{\mathbf{x}}$ is fed to the input layer of the auto-encoder. That is, instead of simply copying the input, the denoising auto-encoder must undo this corruption, thereby implicitly learning useful features from the internal structure of data. In order to limit over-fitting, we have introduced a L1 regularization penalty term α_{act} on the node activities and a L2 regularization penalty term α_{weight} on the weight vector $\mathbf{w}_i$. The objective function thus becomes,

$$loss(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2 + \sum_{i=1}^{l} (\alpha_{act} \ \|F_{1->i}(\tilde{\mathbf{x}})\|^2 + \alpha_{weight} \ \|\mathbf{w}_i\|), \tag{1}$$

where l is the number of layers in the auto-encoder and $\mathbf{z}$ is the output of the auto-encoder taking $\tilde{\mathbf{x}}$ as its input. It is to be noted that $\tilde{\mathbf{x}}$ is the corrupted version of given input sample $\mathbf{x}$. $F_{1->l}(\tilde{\mathbf{x}}) = f_l \circ f_{l-1} \circ \cdots \circ f_1$ is the composed function that defines the the reconstruction $\mathbf{x}'$ for an auto-encoder with l layers and $f_i(x) = \ max \ (0, x_s)$ where x_s is the weighted sum of inputs to a neuron following McCulloch-Pitts model of neuron.

The DDAE network consists of an input layer, an output layer and two auto-encoders in between, with encoding dimensions 500 and 200 respectively. The input layer takes the pre-processed DNAm and CNV combined together as input data, the output layer produces a reconstruction of the input. The values of α_{act} and α_{weight} have been set to 0.0001 and 0.001 respectively. The DDAE has been trained for 25 epochs with a stochastic gradient descent algorithm using Mean Squared Error (MSE) as the loss function.

2.3.1.2. Multi-output regression using multilayer perceptron. Once the DDAE network is trained, the reduced (m) features have been extracted from the bottleneck layer (here, $m = 200$). These features have been fed as input to the MLP-based regression model.

An MLP is a feedforward artificial neural network consisting of an input layer, one or more hidden layers and an output layer. Each layer of neurons (nodes) is fully connected with the next layer of neurons. The nodes in the hidden layer and the output layer are driven by a non-linear activation function. It is a supervised algorithm that maps an input $\mathbf{x} \varepsilon \Re^n$ to an output $\mathbf{z} \varepsilon \Re^m$, where, n is the input dimension and m is the output dimension. Given, a target $\mathbf{t}$ for a set of features, the multilayer perceptron can learn a non-linear estimator for regression.

In the model, the output layer of the regression model has the same number of neurons as the number of target variables in the RNA-seq data. We have used ReLU activation function in the intermediate layers. The output layer of the regression model uses linear activations. The backpropagation algorithm [47] has been used to train the MLP.

Let the error in the i^{th} output node for the training data be denoted by $error_i = t_i - o_i$, where t_i is the target value and o_i is the actual output of the i^{th} neuron. The error in the entire output is thus given by:

$$E = \frac{1}{2} \sum_i error_i^2 \tag{2}$$

The node weights are then corrected to minimize the error using the gradient descent rule. We have modelled the task of estimating gene expression from DNA methylation and copy number variation data as a multi-output regression (also known as multi-target or multi-variate regression) problem, where we have predicted multiple real-valued responses simultaneously. The features extracted from the bottleneck layer of the DDAE network have been used as the training dataset D having s instances $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_s$, where each instance $\mathbf{x}_i$ is characterized by an input vector of k descriptive variables $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{ik}]^T$ and an output vector of t target variables $\mathbf{t}_i = [t_{i1}, t_{i2} \cdots, t_{it}]^T$, k being the number of extracted features (here, m) extracted from the input data and t being the number of genes in the RNA-seq data respectively.

The MLP-regression model uses two hidden layers and has been trained with an 'adam' optimizer for 50 epochs using MSE as the loss function. Overfitting in a deep neural network can be prevented using dropout, a mechanism to randomly drop off some units (neurons) in the visible and hidden layers. We have used a dropout of 20% at the input layer and 50% at the hidden layers as suggested in literature [48]. The results obtained using no dropout, 20% dropout in all layers, and using 20% dropout at the input layer and 50% dropout at the hidden layer have been benchmarked against other classical regression methods. This is provided in Section 3.

2.3.2. Other methods for comparison

In order to compare the results obtained from the proposed DDAE-MLP model, we have chosen six benchmark regression models, based on Linear regression, Lasso, Ridge, Random forest (RF), k-Nearest Neighbors (k-NN) and Support Vector Regression (SVR). Lasso linear model uses a L1 penalty term for regularization and adds sparsity to the coefficients. Ridge regularization technique, on the other hand, uses a L2 penalty term on the size of the coefficients [49]. Both these models are used often interchangeably as prediction models, and are capable of making estimations that are closely correlated with true values. Random forest (also known as Random decision forest) [50] is an ensemble method used for learning with a regression or classification outlook and is known to produce good prediction accuracy. k-NN is another non-paramteric method in which estimations are done considering the k-closest training examples [51]. SVR [52], on the other hand, uses Support Vector Machines for regression and is often used as a benchmark method. Therefore, we have compared results obtained by DDAE-MLP with that of these six models.

To build a robust model, we have divided the dataset into two partitions: train and test. The train dataset has been used for training the model, while the test dataset has been used to evaluate the trained model. To evaluate and compare results from different models, we have used *MSE* and R^2 values. The *MSE* for a prediction model is given by:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2, \tag{3}$$

where y_i and $\tilde{y}_i$ are the observed and the predicted values respectively, and n is the number of samples. The R^2 statistic (also known as coefficient of determination), provides a measure of fit. It gives a ratio of variance explained, and hence it always has a value between 0 and 1. Most importantly, it is independent of the scale of Y.

$$RSS = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2, \tag{4}$$

where RSS is the residual sum of squares.

To calculate R^2, we use the formula:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{5}$$

where $TSS = \sum (y_i - \bar{y})^2$ is the total sum of squares, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ being the mean of the observed data.

2.3.3. Evaluating DDAE extracted features by supervised classification

An auto-encoder usually extracts significant features by minimizing the reconstruction loss. The effectiveness of these extracted features needs to be evaluated. Studies show that measuring the reconstruction loss for data can be an effective way to evaluate auto-encoder-extracted features for unsupervised learning problems, whereas, computing the classification accuracy can be used for supervised learning problems [53,54]. Consequently, in this experiment, the DDAE-extracted features have been evaluated by training a classifier on these features and then validating the model on test set. The classifier used here is a MLP-based classifier. A train-test split of 80 − 20% has been used. In the test set, 69 samples were from cancer patients and 12 of them were from healthy patients. The features extracted from the bottleneck layer of the DDAE network have been used as the input to the MLP-based classifier. The classification performed on the DDAE-extracted features has shown promising results as shown in Section 3.

3. Results

This section shows results obtained from the proposed regression model for estimating Gene Expression values from DNA Methylation and Copy Number Variation data against those obtained from benchmark regression methods.

3.1. Estimating gene expression values from DNA methylation and copy number variation profiles

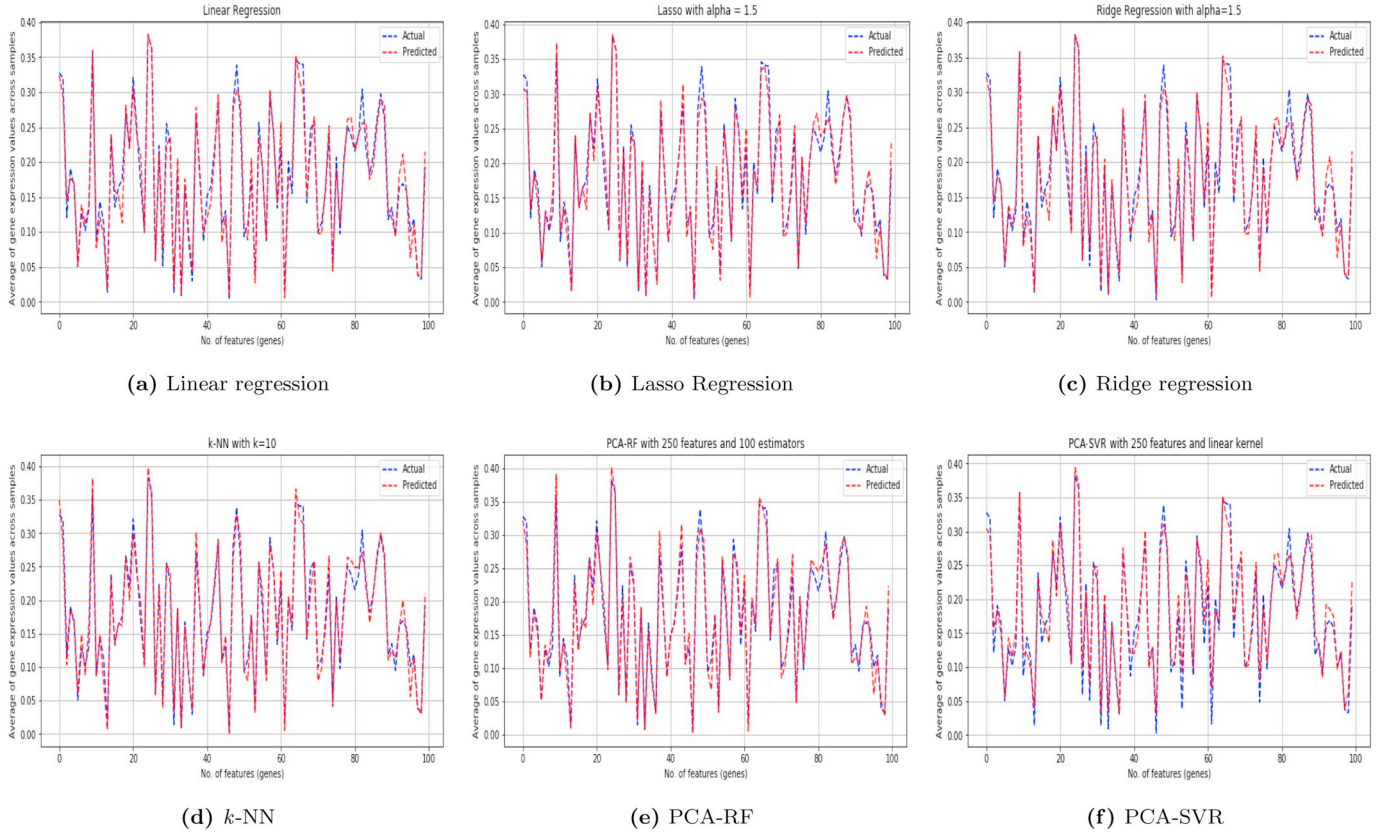
In this work, we have developed a deep learning-based regression model to predict RNA-seq values from DNA methylation and copy number variation profiles. The features extracted from the integrated DNA methylation and copy number variation data form the input to the regression model. The plot for true gene expression values vs. predicted gene expression values across all samples is shown in Fig. S5 (in Supplementary Material).

Fig. 2 shows a zoomed view, a plot for 100 randomly selected genes obtained using the benchmark regression methods, whereas, Fig. 3 shows similar plot for the proposed DDAE-MLP based method. We have observed that actual and predicted values show similar up-down regulation patterns for genes, showing similar peak and trough points. This is significant in cases where exact gene expression values are not known and only up-down regulation trends are to be compared. The average of true gene expression values across all samples also show a positive correlation with the average of predicted gene expression values across all samples, with a R^2 value of 0.968 and a correlation coefficient of 0.983, as illustrated in Fig. S6 (in Supplementary Material).

When compared to other standard regression models based on Linear regression, Lasso, Ridge, k-NN, RF and SVR, the proposed DDAE-MLP model has shown either better or closely comparable results. In order to reduce computational complexity, PCA has been used first, followed by RF and SVR separately. Experiments to estimate MSE and R^2 values with different hyper-parameter settings for the penalized regression techniques, have been conducted. The results have been compared with that obtained using the proposed DDAE-MLP model both with and without dropout. The MSE plot for the DDAE-MLP model compared to all other benchmark models, ranked on negative logarithmic scale of MSE, has been shown in Fig. 4. The proposed DDAE-MLP model, with 20% dropout has topped the chart, followed by Ridge and other methods. Fig. 5 shows the geom-bar graph comparing the R^2 values obtained by different methods and the DDAE-MLP method produces an R^2 value very close to the best result.

3.2. Classification of tumor vs. normal samples using DDAE extracted features

We have further tested the effectiveness of the features extracted from DNA methylation and copy number variation data by performing



(a) Linear regression **(b)** Lasso Regression **(c)** Ridge regression

(d) *k*-NN **(e)** PCA-RF **(f)** PCA-SVR

Fig. 2. Actual and predicted gene expression values for 100 randomly selected genes for benchmark methods.

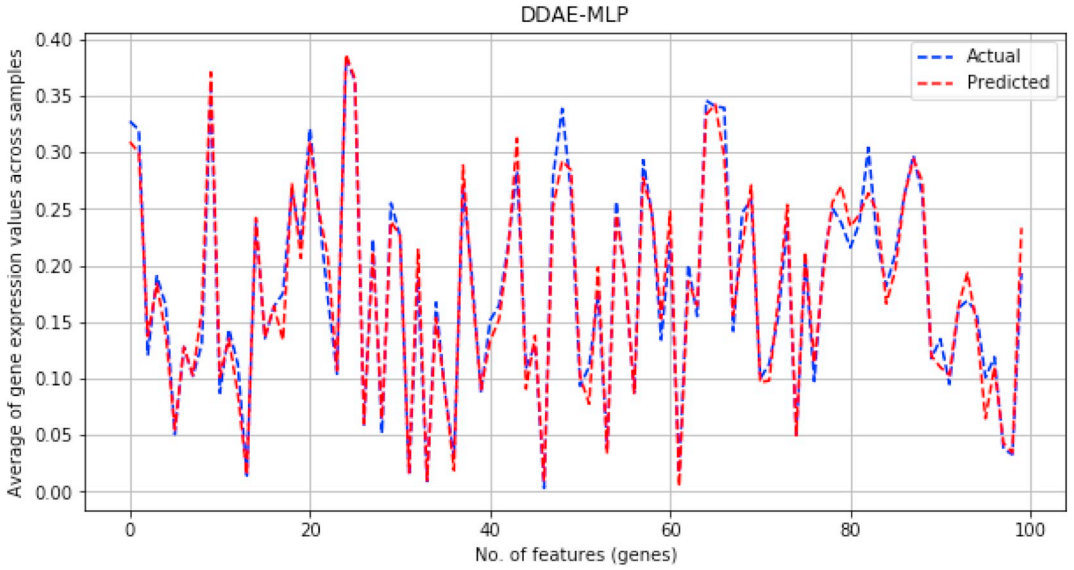


Fig. 3. Actual and predicted gene expression values for 100 randomly selected genes for the proposed DDAE-MLP method.

classification of tumor and normal samples. Among 404 samples, the number of normal and cancer samples were 45 and 359 respectively. A train-test split of 80 − 20% has been used on the TCGA data before they have been fed into the DDAE network. Out of 81 samples in the test dataset, 69 were cancer and 12 were normal. The features extracted by the DDAE have shown good classification performance with an accuracy of 95.1%. Here, 12 out of 12 normal samples and 65 out of 69 tumor samples have been classified correctly. The precision, recall and F1-score values have been found to be 0.96, 0.95 and 0.95 respectively.

4. Discussion and conclusion

In this work, we have explored the impact of both genomic and epigenomic features on gene expression regulation using a deep learning-based regression model. The model has established the fact that the integrated features can be used to predict gene expression patterns and also serves as a promising platform for multi-omics integration. The model also captures the directional regulation of the predicted gene expression.

TCGA provides us with tremendous resources of high-quality cancer molecular data but we are still often limited by sample size while using

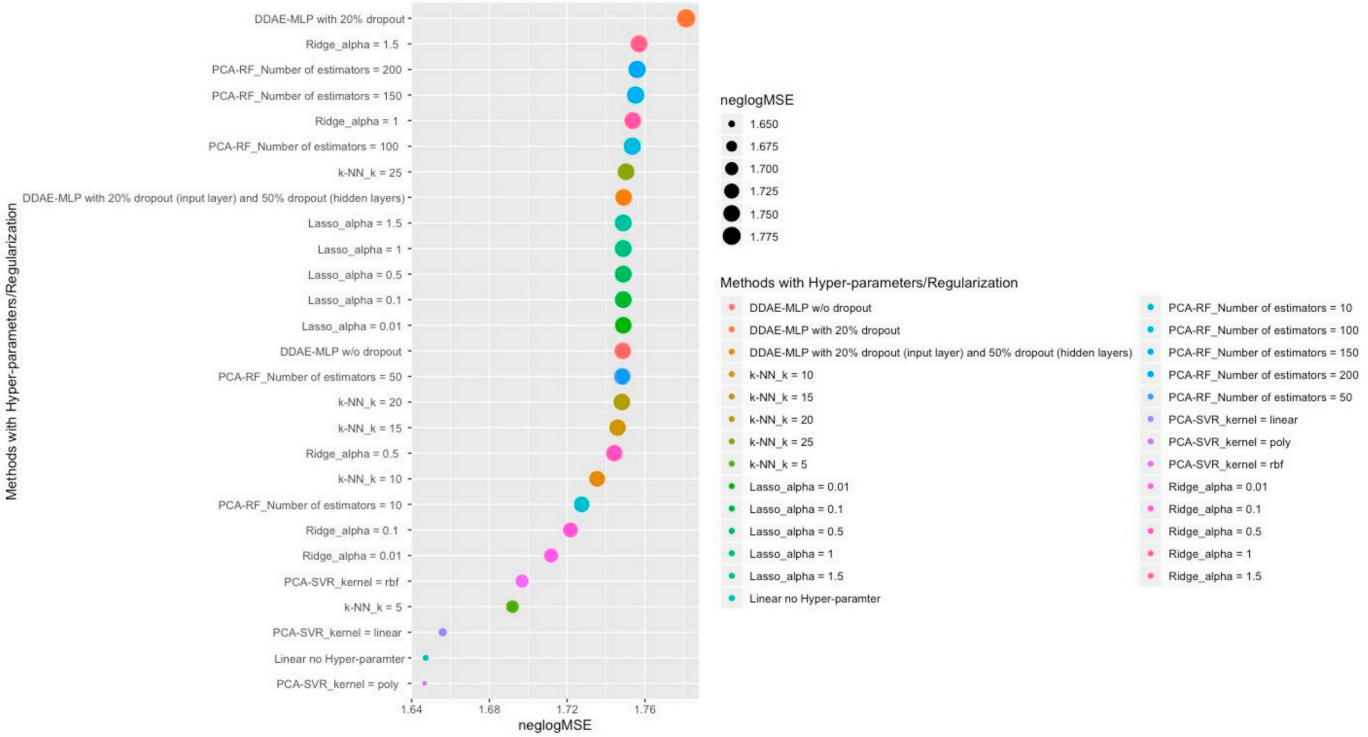


Fig. 4. MSE plot for the DDAE-MLP method compared with other benchmark methods, ranked based on neglog(MSE).

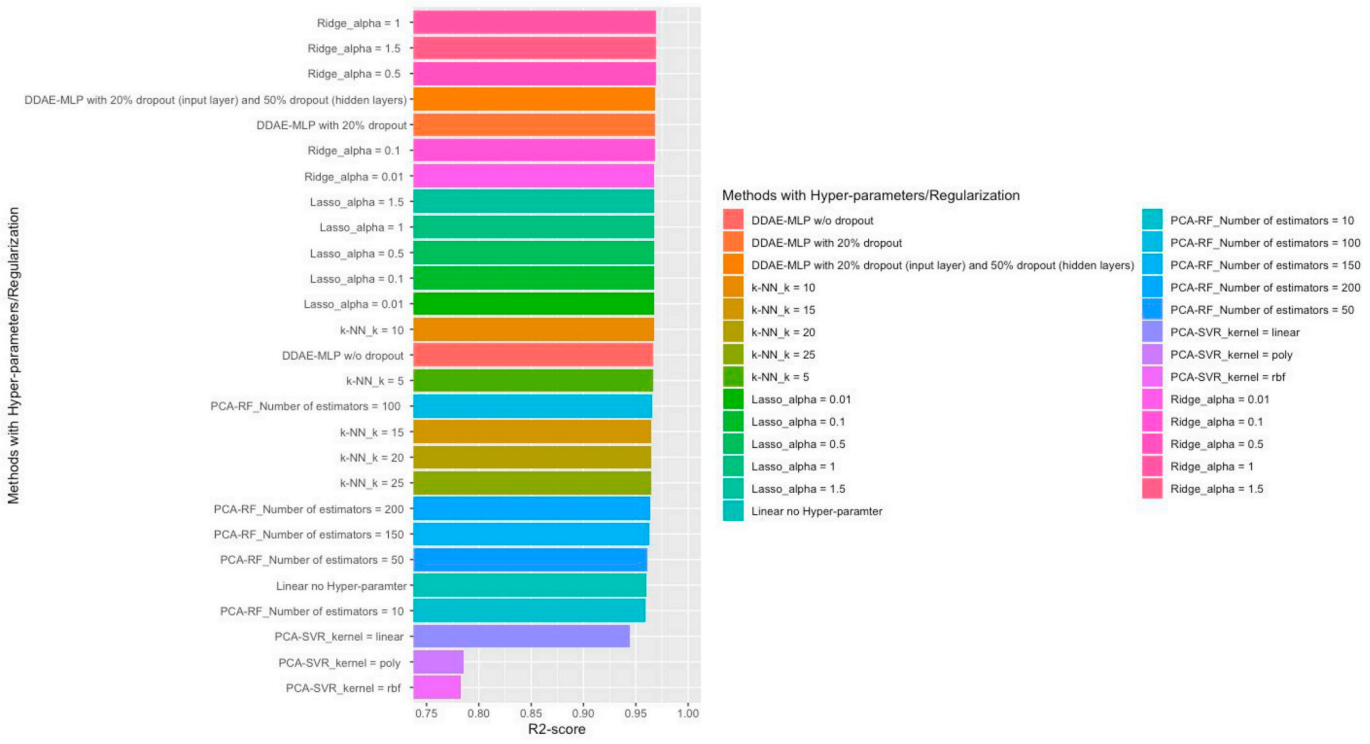


Fig. 5. R^2 value for DDAE-MLP method compared with other benchmark methods, ranked based on R^2 value.

deep learning based strategies for multi-omics integration. In this current scope of work, we have used 404 matched samples having all three DNAm, CNV and RNA-seq. A lot can be leveraged if we can use samples from archival tissue BioBanks like formalin-fixed (FF) and paraffin-embedded (FFPE) samples. This enables us to increase the sample size and also add more input features to the model. In future, we would like to extend this work on clinical samples from Tissue BioBanks having genomic (CNV) and epigenomic (DNAm) in order to estimate gene expression by comparing directly with true gene expression data available in those paired samples. In the case of denatured mRNA or no availability, estimated gene expression from DNAm and CNV can be correlated with high-quality TCGA mRNA data. Multiple literature evidence has often pointed out the potential limitations for usage of FF/FFPE datasets due to sampling quality/degradation of material associated as outlined in [55,56] leading to poor quality gene-expression readouts, limiting downstream analysis or any mechanistic understanding.

Our current DDAE-MLP based regression method also accounts for stochasticity, nonlinear properties, signal-to-noise, heterogeneity enough from the given input features to infer gene expression close to its true values. This model can be useful if there are large CNV and/or methylation data for patients with tumor where we cannot have gene expression values since the mRNA data are not available or not of high quality. We could still estimate gene expression values that would be enough to classify diseases, their progression and mechanism. This would let us tackle issues arising out of poor or low quality mRNA data in the process. Thus it provides us with an advantage to design and test future integrative multi-omics, multi-platform studies using datasets from multiple cohorts for patients with genomic and epigenomic data. This would not only increase our pool of matched samples but also provide us with a better mechanistic characterization of the disease. Thus, DDAE-MLP has been able to find gene expression surrogates derived from DNAm and CNV without having real RNA-seq expression, and has produced better, in some cases, at least as good a result as obtained by benchmark regression methods. The features extracted from DNAm and CNV have also led to good tumor/normal classification performance.

As an extension of the work, we have also tried to estimate CNV as a function of DNAm, in a separate experiment. The methodology used for estimating CNV and the corresponding results have been explained in Section S1 and Section S2 (in Supplementary Material) respectively.

Currently there are some limitations in this study. One of the limitations is the absence of the compute time and infrastructure resource comparison used in running all the models during the benchmark process. In the current scope of work, our promoter regions are confined within 1500 bps up and downstream of TSS. All the CpG islands are also confined within this window-size. CNVs can often extend beyond this region. This precise selection window could be a second limitation. We have not used any synthetic simulation data as a surrogate data-type for any training or test purposes. This could potentially be a third limitation. However, our primary goal was to capture realistic non-linear features from the real multi-omics data-sets and benchmark the various regression models which is often missed in synthetic data creation based on pre-specified metrics. Hence, we have discounted the usage of synthetic datasets in this scope of work. We are also currently limited to only single cancer type at our end in this scope of work. We would like to extend the utility of this work across other cancer types in future, to evaluate the robustness of the proposed DDAE-MLP method. We would also like to extend the future investigations to delineate top predicted gene expression features and molecular pathways from the same. This could aid in identification of molecular sub-stratification of patient cohorts. This work can also be extended to study the effect of histone modifications on gene expressions. All these integration taken together would thus serve as a potential platform to better delineate the gene regulatory framework of the disease and enhance our understanding of the phenotype.

Data and Code Availability

The source codes have been implemented in Python 3 and are freely available at https://github.com/vd4mmind/multiOmicsIntegration. The data used in this study can be downloaded from https://zenodo.org/record/3712496#.XnBlS5NKjGI.

Acknowledgement

DBS thanks the Data Science Laboratory, A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India for providing the computing infrastructure to execute and test some of the models. VD currently works as a Post Doctoral researcher in Novo Nordisk. However, he did not receive any funding for this work. RKD acknowledges the financial support received from the inter-institutional Systems Medicine Cluster project (BT/Med-II/NIBMG/SyMeC/2014/Vol. II) from the Department of Biotechnology, Government of India.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.03.021.

References

[1] J.S. You, P.A. Jones, Cancer genetics and epigenetics: Two sides of the same coin? Cancer Cell 22 (1) (2012) 9–20 22789535[pmid] https://doi.org/10.1016/j.ccr.2012.06.008 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396881/.

[2] N.-A.-D. Marzouka, J. Nordlund, C.L. Bäcklin, G. Lönnerholm, A.-C. Syvänen, J. Carlsson Almlöf, Copynumber 450k cancer: baseline correction for accurate copy number calling from the 450k methylation array, Bioinformatics (Oxford, England) 32 (7) (2016) 1080–1082 26553913[pmid] https://doi.org/10.1093/bioinformatics/btv652 https://www.ncbi.nlm.nih.gov/pubmed/26553913.

[3] A. Feber, P. Guilhamon, M. Lechner, T. Fenton, G.A. Wilson, C. Thirlwell, T.J. Morris, A.M. Flanagan, A.E. Teschendorff, J.D. Kelly, S. Beck, Using high-density dna methylation arrays to profile copy number alterations, Genome Biol. 15 (2) (2014) R30, https://doi.org/10.1186/gb-2014-15-2-r30. URL http://europepmc.org/articles/PMC4054098.

[4] D.H. Lim, E.R. Maher, DNA methylation: A form of epigenetic control of gene expression, Obstetr. Gynaecol. 12 (1) (2010) 37–42.

[5] A. Shlien, D. Malkin, Copy number variations and cancer, Genome Med. 1 (6) (2009) 62 gm62[PII] https://doi.org/10.1186/gm62 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703871/.

[6] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavaré, P. Deloukas, M.E. Hurles, E.T. Dermitzakis, Relative impact of nucleotide and copy number variation on gene expression phenotypes, Science 315 (5813) (2007) 848–853 https://science.sciencemag.org/content/315/5813/848.full.pdf https://doi.org/10.1126/science.1136678.

[7] C. Zhou, W. Zhang, W. Chen, Y. Yin, M. Atyah, S. Liu, L. Guo, Y. Shi, Q. Ye, Q. Dong, N. Ren, Integrated analysis of copy number variations and gene expression profiling in hepatocellular carcinoma, Sci. Rep. 7 (1) (2017) 10570, https://doi.org/10.1038/s41598-017-11029-y.

[8] J. Shen, S. Wang, Y.-J. Zhang, H.-C. Wu, M.G. Kibriya, F. Jasmine, H. Ahsan, D.P. Wu, A.B. Siegel, H. Remotti, R.M. Santella, Exploring genome-wide dna methylation profiles altered in hepatocellular carcinoma using infinium humanmethylation 450 beadchips, Epigenetics 8 (1) (2013) 34–43 2012EPI0258R [PII] https://doi.org/10.4161/epi.23062 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549879/.

[9] R.A. Hlady, D. Zhou, W. Puszyk, L.R. Roberts, C. Liu, K.D. Robertson, Initiation of aberrant dna methylation patterns and heterogeneity in precancerous lesions of human hepatocellular cancer, Epigenetics 12 (3) (2017) 215–225 pMID: 28059585 https://doi.org/10.1080/15592294.2016.1277297.

[10] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–74, https://doi.org/10.1038/nature11247 http://europepmc.org/articles/PMC3439153.

[11] A. Ally, et al., Comprehensive and integrative genomic characterization of hepatocellular carcinoma, Cell 169 (7) (2017) 1327–1341.e23, https://doi.org/10.1016/j.cell.2017.05.046.

[12] M. Kim, N. Rai, V. Zorraquino, I. Tagkopoulos, Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*, Nat. Commun. 7 (1) (2016) 13090, https://doi.org/10.1038/ncomms13090.

[13] M.-S. Kwon, Y. Kim, S. Lee, J. Namkung, T. Yun, S.G. Yi, S. Han, M. Kang, S.W. Kim, J.-Y. Jang, T. Park, Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer, BMC Genomics 16 (Suppl 9) (2015) S4 1471--2164-16-S9-S4[PII] https://doi.org/10.1186/1471-2164-16-S9-S4 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547403/.

[14] S. Pineda, F.X. Real, M. Kogevinas, A. Carrato, S.J. Chanock, N. Malats, K. Van Steen, Integration analysis of three omics data using penalized regression methods: An application to bladder cancer, PLoS Genet. 11 (12) (2015) 1–22, https://doi.org/10.1371/journal.pgen.1005689.

[15] Y. Zhang, A. Li, J. He, M. Wang, A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data, IEEE J. Biomed. Health Inform. 24 (1) (2020) 171–179, https://doi.org/10.1109/JBHI.2019.2898471.

[16] D. Sun, A. Li, B. Tang, M. Wang, Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome, Comput. Methods Programs Biomed. 161. https://doi.org/10.1016/j.cmpb.2018.04.008.

[17] D. Wang, J. Gu, Integrative clustering methods of multi-omics data for molecule-based cancer classifications, Quant. Biol. 4 (1) (2016) 58–67, https://doi.org/10.1007/s40484-016-0063-4.

[18] M. Kim, I. Tagkopoulos, Data integration and predictive modeling methods for multi-omics datasets, Mol. Omics 14 (2018) 8–25, https://doi.org/10.1039/C7MO00051K.

[19] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, K.-A. Lê Cao, Diablo: From multi-omics assays to biomarker discovery, an integrative approach, bioRxiv https://www.biorxiv.org/content/early/2018/03/20/067611.full.pdf, https://doi.org/10.1101/067611.

[20] A. Ahmad, H. Fröhlich, Integrating heterogeneous omics data via statistical inference and learning techniques, Genom. Computat. Biol. 2 (1) (2016) e32, https://doi.org/10.18547/gcb.2016.vol2.iss1.e32 https://genomicscomputbiol.org/ojs3/GCB/article/view/28.

[21] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, L. Milanesi, Methods for the integration of multi-omics data: Mathematical aspects, BMC Bioinform. 17 (2) (2016) S15, https://doi.org/10.1186/s12859-015-0857-9.

[22] K. Chaudhary, O.B. Poirion, L. Lu, L.X. Garmire, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clin. Cancer Res. 24 (6) (2018) 1248–1259 https://clincancerres.aacrjournals.org/content/24/6/1248.full.pdf https://doi.org/10.1158/1078-0432.CCR-17-0853.

[23] O.B. Poirion, K. Chaudhary, L.X. Garmire, Deep learning data integration for better risk stratification models of bladder cancer, AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science 2017, 2018, pp. 197–206 29888072[pmid] https://www.ncbi.nlm.nih.gov/pubmed/29888072.

[24] K. Chaudhary, O.B. Poirion, L. Lu, S. Huang, T. Ching, L.X. Garmire, Multimodal meta-analysis of 1,494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes, Clin. Cancer Res. 25 (2) (2019) 463–472 https://clincancerres.aacrjournals.org/content/25/2/463.full.pdf https://doi.org/10.1158/1078-0432.CCR-18-0088.

[25] D. Sun, M. Wang, A. Li, A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data, IEEE/ACM Trans. Computat. Biol. Bioinform. 16 (3) (2019) 841–850, https://doi.org/10.1109/TCBB.2018.2806438.

[26] H. Kazan, Modeling Gene Regulation in Liver Hepatocellular Carcinoma with Random Forests 2016, (2016), p. 6, https://doi.org/10.1155/2016/1035945.

[27] Y. Chen, Y. Li, R. Narayan, A. Subramanian, X. Xie, Gene expression inference with deep learning, Bioinformatics 32 (12) (2016) 1832–1839, https://doi.org/10.1093/bioinformatics/btw074.

[28] J. Li, T. Ching, S. Huang, L.X. Garmire, Using epigenomics data to predict gene expression in lung cancer, BMC Bioinform. 16 (5) (2015) S10, https://doi.org/10.1186/1471-2105-16-S5-S10.

[29] R. Xie, J. Wen, A. Quitadamo, J. Cheng, X. Shi, A deep auto-encoder model for gene expression prediction, BMC Genomics 18 (9) (2017) 845, https://doi.org/10.1186/s12864-017-4226-0.

[30] H.K. Solvang, O.C. Lingjærde, A. Frigessi, A.-L. Børresen-Dale, V.N. Kristensen, Linear and non-linear dependencies between copy number aberrations and mrna expression reveal distinct molecular pathways in breast cancer, BMC Bioinform. 12 (2011) 197 21609452[pmid] https://doi.org/10.1186/1471-2105-12-197 https://www.ncbi.nlm.nih.gov/pubmed/21609452.

[31] G. Lenka, M.-H. Tsai, H.-C. Lin, J.-H. Hsiao, Y.-C. Lee, T.-P. Lu, J.-M. Lee, C.-P. Hsu, L.-C. Lai, E.Y. Chuang, Identification of methylation-driven, differentially expressed stxbp6 as a novel biomarker in lung adenocarcinoma, Sci. Rep. 7 (2017) 42573, https://doi.org/10.1038/srep42573 http://europepmc.org/articles/PMC5309775.

[32] G. Lee, L. Bang, S.Y. Kim, D. Kim, K.-A. Sohn, Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer, BMC Med. Genom. 10 (Suppl 1) (2017) 28 28589855[pmid] https://doi.org/10.1186/s12920-017-0268-z https://www.ncbi.nlm.nih.gov/pubmed/28589855.

[33] X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, X. Fan, Copy number variation is highly correlated with differential gene expression: a pan-cancer study, BMC Med. Genet. 20 (1) (2019) 175, https://doi.org/10.1186/s12881-019-0909-5.

[34] H. Zhong, S. Kim, D. Zhi, X. Cui, Predicting gene expression using DNA methylation in three human populations, PeerJ 7 (2019) e6757 31106051[pmid] https://doi.org/10.7717/peerj.6757 https://pubmed.ncbi.nlm.nih.gov/31106051.

[35] O. Gevaert, R. Tibshirani, S.K. Plevritis, Pancancer analysis of dna methylation-driven genes using methylmix, Genome Biol. 16 (1) (2015) 17, https://doi.org/10.1186/s13059-014-0579-8.

[36] J. Tan, M. Ung, C. Cheng, C.S. Greene, Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders, Pacific Symposium on Biocomputing 20, 2015, pp. 132–143 25592575[pmid] https://www.ncbi.nlm.nih.gov/pubmed/25592575.

[37] P. Danaee, R. Ghaeini, D.A. Hendrix, A deep learning approach for cancer detection and relevant gene identification, Pacific Symposium on Biocomputing 22, 2017, pp. 219–229 27896977[pmid] https://doi.org/10.1142/9789813207813_0022 https://www.ncbi.nlm.nih.gov/pubmed/27896977.

[38] F.M. Alakwaa, K. Chaudhary, L.X. Garmire, Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data, J. Proteome Res. 17 (1) (2018) 337–347 29110491[pmid] https://doi.org/10.1021/acs.jproteome.7b00595 https://www.ncbi.nlm.nih.gov/pubmed/29110491.

[39] L. Wei, Z. Jin, S. Yang, Y. Xu, Y. Zhu, Y. Ji, Tcga-assembler 2: Software pipeline for retrieval and processing of tcga/cptac data, Bioinformatics (Oxford, England) 34 (9) (2018) 1615–1617 29272348[pmid] https://doi.org/10.1093/bioinformatics/

btx812 https://www.ncbi.nlm.nih.gov/pubmed/29272348.
[40] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (3) (2014) 333.
[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830 http://dl.acm.org/citation.cfm?id=1953048.2078195.
[42] F. Martínez-Murcia, A. Ortiz, J. Gorriz, J. Ramírez, D. Castillo-Barnes, D. Salas-Gonzalez, F. Segovia, Deep convolutional autoencoders vs pca in a highly-unbalanced parkinson's disease dataset: A datscan study, (2019), pp. 47–56, https://doi.org/10.1007/978-3-319-94120-2_5.
[43] M. Sakurada, T. Yairi, Anomaly detection using autoencoders with nonlinear dimensionality reduction, Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis, MLSDA'14, ACM, New York, NY, USA, 2014, pp. 4:4–4:11, , https://doi.org/10.1145/2689746.2689747.
[44] Y. Wang, H. Yao, S. Zhao, Auto-encoder based dimensionality reduction, Neurocomput. 184 (C) (2016) 232–242, https://doi.org/10.1016/j.neucom.2015.08.104.
[45] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Parallel Distributed Processing – Explorations in the Microstructure of Cognition, MIT Press, 1986, pp. 318–362 Ch. 8.
[46] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, Proceedings of the 25th International Conference on Machine Learning, ICML '08, ACM, New York, NY, USA, 2008, pp. 1096–1103, , https://doi.org/10.1145/1390156.1390294.
[47] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536, https://doi.org/10.1038/323533a0 http://www.nature.com/articles/323533a0.
[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958 http://dl.acm.org/citation.cfm?id=2627435.2670313.
[49] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc, New York, NY, USA, 2001.
[50] T.K. Ho, Random decision forests, Proceedings of the Third International Conference on Document Analysis and Recognition - Volume 1, ICDAR '95, IEEE Computer Society, Washington, DC, USA, 1995, p. 278 http://dl.acm.org/citation.cfm?id=844379.844681.
[51] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (2006) 21–27, https://doi.org/10.1109/TIT.1967.1053964.
[52] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Advances in Neural Information Processing Systems 9, MIT Press, 1997, pp. 155–161.
[53] Q. Meng, D. Catchpoole, D. Skillicom, P.J. Kennedy, Relational autoencoder for feature extraction, 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 364–371, , https://doi.org/10.1109/IJCNN.2017.7965877.
[54] M.F. Baln, A. Abid, J. Zou, Concrete autoencoders: Differentiable feature selection and reconstruction, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Vol. 97 of Proceedings of Machine Learning Research, PMLR, Long Beach, California, USA, 2019, pp. 444–453 http://proceedings.mlr.press/v97/balin19a.html.
[55] P.P. Reis, L. Waldron, R.S. Goswami, W. Xu, Y. Xuan, B. Perez-Ordonez, P. Gullane, J. Irish, I. Jurisica, S. Kamel-Reid, mRNA transcript quantification in archival samples using multiplexed, color-coded probes, BMC Biotechnol. 11 (2011) 46 21549012[pmid] https://doi.org/10.1186/1472-6750-11-46 https://www.ncbi.nlm.nih.gov/pubmed/21549012.
[56] L.N. Kwong, M.P. De Macedo, L. Haydu, A.Y. Joon, M.T. Tetzlaff, T.L. Calderone, C.-J. Wu, M.K. Kwong, J. Roszik, K.R. Hess, M.A. Davies, A.J. Lazar, J.E. Gershenwald, Biological validation of rna sequencing data from formalin-fixed paraffin-embedded primary melanomas, JCO Precision Oncol. 2018 (2018), https://doi.org/10.1200/PO.17.00259 31058252[pmid] https://www.ncbi.nlm.nih.gov/pubmed/31058252.