Learning data representation using modified autoencoder for the integrative analysis of multi-omics data

Tianwei Yu

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA.

Abstract

In integrative analyses of omics data, it is often of interest to extract data embedding from one data type that best reflect relations with another data type. This task is traditionally fulfilled by linear methods such as canonical correlation and partial least squares. However, information contained in one data type pertaining to the other data type may not be in the linear form. Deep learning provides a convenient alternative to extract nonlinear information. Here we develop a method Autoencoder-based Integrative Multi-omics data Embedding (AIME) to extract such information. Using a real gene expression – methylation dataset, we show that AIME extracted meaningful information that the linear approach could not find. The R implementation is available at http://web1.sph.emory.edu/users/tyu8/AIME/.

1 Introduction

It is becoming more and more common that multiple omics data are collected on the same set of subjects to obtain a global view of the molecular signature of a disease. When analyzing such data, a common task is to find data embedding in a lower dimensional space from one data type that best preserves the information pertaining to another data type. Current approaches include Canonical Correlation Analysis (CCA), Partial Least Squares (PLS) and their variants, which are based on (sparse) linear projection of the data (Safo, et al., 2018). Given the complexity of omics data, a nonlinear equivalent is desired but is currently unavailable.

Autoencoder is a deep learning – based nonlinear embedding approach that is typically used to achieve sparse data representation from a single dataset (Tan, et al., 2015), reduce noise (Eraslan, et al., 2019), impute missing values (Talwar, et al., 2018), conduct pre-training for classification tasks (Pan, et al., 2016), and make functional inferences (Danaee, et al., 2017; Gligorijevic, et al., 2018). In terms of
integrative analysis, Chaudhary et al. (2018) used the original autoencoder structure and the combined multi-omics data in both the input and reconstruction layers in order to find their interactions for the prediction task.

In this manuscript, we modify the autoencoder structure by using two data types in the input and reconstruction layers respectively, in order to extract nonlinear data embedding from one data type that is associated with the other data type. This work can be seen as a nonlinear equivalence to canonical correlation analysis.

2 Methods

We set up a structure that is similar to autoencoder, as shown in Fig. 1a. Different from the typical autoencoder, the input layer and reconstruction layer use two different omics data types. Thus in some sense this is a prediction structure with very high dimensional outcome. Such a prediction task is unrealistic, and our goal is not prediction. With a very narrow bottleneck layer in the middle, we essentially seek a nonlinear dimension reduction of the input data type, which best preserves the information pertaining to the output data type.

In the current proof-of-concept study, we used a structure with 3 hidden layers with dropout in the encoder, and 3 hidden layers with dropout in the decoder. Suppose there are \( p \) input variables and \( q \) output variables, the three encoder layers contained \( p/5 \) (with 20% dropoff), \( p/25 \) (with 10% dropoff), and \( p/625 \) nodes respectively. Similarly, the three output decoder layers contained \( q/625 \), \( q/25 \) (with 10% dropoff) and \( q/5 \) (with 20% dropoff) nodes respectively. At the center, the number of nodes in the bottleneck layer depends on the number of dimensions the user desires, which was set at 4 in the illustration study.

In order to select top contributing variables, our main focus is their influence on data embedding. After model fitting, keeping the fitted weights, we iteratively permute each input variable 10 times, and
compute the average sum of squares difference of the embedded points to the original ones. The variables resulting in the largest sum of squared shifts are considered most influential.

![Diagram](image)

**Figure 1.** The model setup and an example application. (a) The setup of the model. (b) An example data embedding result using the MESA Epigenomics and Transcriptomics data (GSE56046/47). Data points are colored based on race/gender/site attributes of the samples. (c) Top 5 GO biological
processes overrepresented by the top 1% important genes, after manual removal of highly overlapping GO terms.

3 Results and Discussion

We illustrate the application using the GSE56046/GSE56047 dataset from the MESA epigenomics and transcriptomics Study, which was conducted on purified human monocytes from a large study population (Reynolds, et al., 2014). We filtered the normalized gene expression data with the criterion of coefficient of variation (CV) > 0.05, and the DNA methylation M-values with the criterion of standard deviation > 1.25. The resulting data contained 5459 genes and 5703 CpG loci in 1202 samples.

For ease of interpretation, we focused on the nonlinear embedding of gene expression data that is associated with DNA methylation. The gene expression data was used as input of the encoder, and the methylation data was used as output. Figure 1b shows the embedded data in four dimensions by AIME (upper right), as compared to canonical correlation analysis (lower left). The data points are colored based on the “racegendersite” variable of the dataset, which is a combined factor coding for the combinations of race, gender, and data collection site. As is clearly demonstrated by the separation of the colored data points, the AIME results indicate the association between gene expression and DNA methylation is strongly dependent on the race and gender, while the CCA results doesn’t show a clear signal. We tested all the other available factors in the data, and no other clear relation to the embedded data pattern was found.

We next examined the top 1% of genes contributing to the date embedding. Among the top 5 GO biological processes found by GOstats (Figure 1c), two were directly related to the cell type and its functionality – “regulation of mononuclear cell proliferation” and “immune system development”. The other three were critical cell development and maintenance processes. The results made clear biological sense.

Overall, AIME extracted novel patterns that are not found by the traditional approach, and the results were easily interpretable. We believe AIME is a valuable addition to the current methods of omics data integrative analyses.

Acknowledgements
The author thanks Mr. Yunchuan Kong and Dr. Hao Wu for helpful discussions. This work has been partially supported by NIH grant R01GM124061.

**References**

Chaudhary, K., *et al.* Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* 2018;24(6):1248-1259.

Danaee, P., Ghaeini, R. and Hendrix, D.A. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. *Pac Symp Biocomput* 2017;22:219-229.

Eraslan, G., *et al.* Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;10(1):390.

Gligorijevic, V., Barot, M. and Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* 2018;34(22):3873-3881.

Pan, X., *et al.* IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics* 2016;17:582.

Reynolds, L.M., *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat Commun* 2014;5:5366.

Safo, S.E., Li, S. and Long, Q. Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. *Biometrics* 2018;74(1):300-312.

Talwar, D., *et al.* AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* 2018;8(1):16329.

Tan, J., *et al.* Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp Biocomput* 2015:132-143.