HetEnc: A Deep Learning Predictive Model for Multi-type Biological Dataset

Leihong Wu, Xiangwen Liu, Weida Tong, Joshua Xu

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Rd., Jefferson, AR 72079, USA

Abstract

Background: Researchers today are generating unprecedented amounts of biological data. One trend in current biological research is integrated analysis with multi-platform data. Effective integration of multi-platform data into the solution of a single or multi-task classification problem; however, is critical and challenging. In this study, we proposed HetEnc, a novel deep learning-based approach, for information domain separation.

Results: HetEnc includes both an unsupervised feature representation module and a supervised neural network module to handle multi-platform gene expression datasets. It first constructs three different encoding networks to represent the original gene expression data using high-level abstracted features. A six-layer fully-connected feed-forward neural network is then trained using these abstracted features for each targeted endpoint. We applied HetEnc to the SEQC neuroblastoma dataset to demonstrate that it outperforms other machine learning approaches. Although we used multi-platform data in feature abstraction and model training, HetEnc does not need multi-platform data for prediction, enabling a broader application of the trained model by reducing the cost of gene expression profiling for new samples to a single platform. Thus, HetEnc provides a new solution to integrated gene expression analysis, accelerating modern biological research.

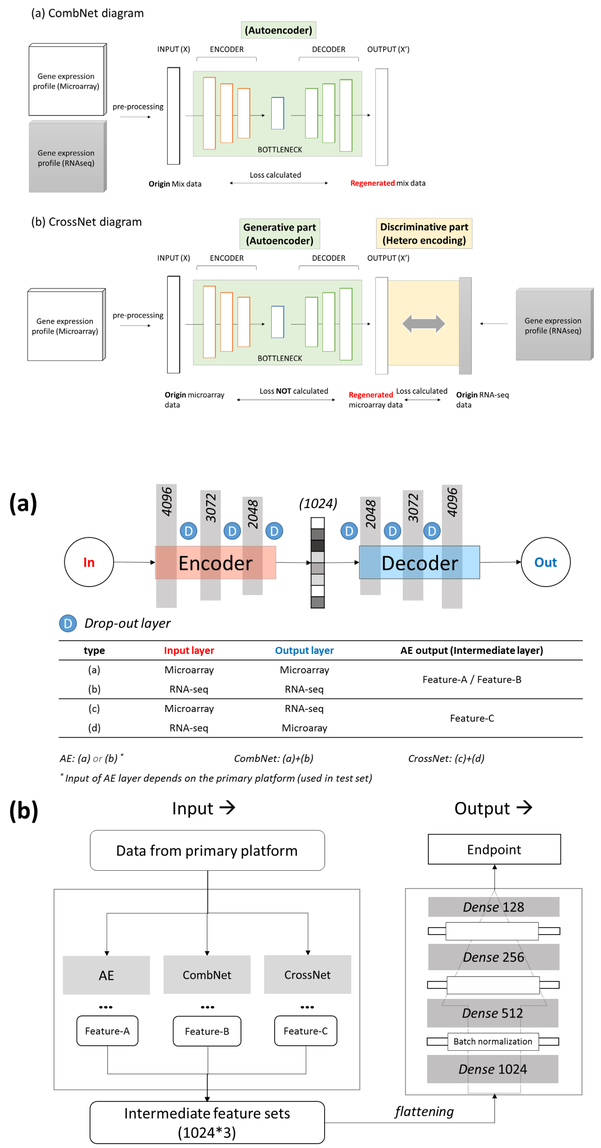

Figure 2. HetEnc overview. (a) feature representation model architecture and three different encoding networks (AE, CombNet and CrossNet) used in the study; (b) feature extraction and 6-DNN structure in the modeling step.

Figure 1. (a) Diagram of CombNet. Microarray and RNA-seq data were mixed before entering the autoencoder. Same feature spaces were defined in both platforms **(b) Diagram of CrossNet.** The first part (generative part) is an autoencoder, where an encoder and decoder are combined to regenerate microarray gene expression profile. The second part (discriminative part) is then introduced to reduce the difference between regenerated microarray data (i.e., the output of generative part) and origin RNA-seq data. In current version, we do not build another discriminative model but use the crossentropy to simplify the process.

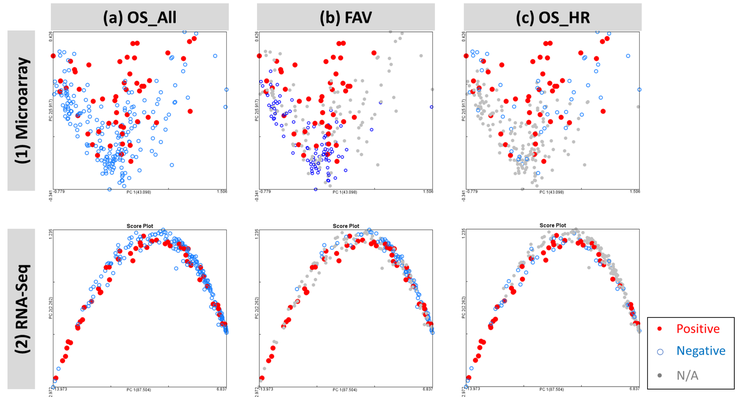

Figure 3. Principle Component Analysis (PCA) by features extracted by AE, CombNet and CrossNet in Microarray data. Intermediate features from AE, CombNet and CrossNet are combined before PCA. By column: (a) 249 training samples labeled by endpoint OS_All; (b) 136 training samples labeled by endpoint FAV; (c) 86 training samples labeled by endpoint OS_HR. By row: input data (1) from Microarray platform; (2) from RNA-seq platform.

Table 1. *Summary of Neuroblastoma Endpoints*

Endpoint	FAV	OS_All	OS_HR
Full description	Neuroblastoma Favorable Prognosis	Overall Survival	Survival in High Risk patients
Sample size (Train/Test)	136/136	249/249	86/90
Train set prevalence	45/91 (0.669)	51/198 (0.795)	43/43 (0.500)
Test set prevalence	46/90 (0.662)	54/195 (0.783)	49/41 (0.544)
Predicting difficulty (Zhang, et al., 2015)	Easy	Medium	Hard

Table 2. *Predictive performance (AUC) for the neuroblastoma dataset*

	Model	RNA-seq			Microarray		
		FAV	OS_All	OS_HR	FAV	OS_All	OS_HR
Cross-validation	**HetEnc**	0.964 (0.009)	0.830 (0.019)	0.520 (0.044)	0.962 (0.011)	0.849 (0.024)	0.651 (0.044)
External Testing (on same testing set)	**HetEnc**	**0.969 (0.007)**	**0.854 (0.024)**	**0.592 (0.027)**	**0.948 (0.015)**	**0.825 (0.016)**	**0.569 (0.022)**
	KNN	0.896 (0.032)	0.641 (0.032)	0.495 (0.048)	0.907 (0.035)	0.662 (0.031)	0.515 (0.041)
	NSC	0.901 (0.036)	0.700 (0.048)	0.499 (0.036)	0.921 (0.032)	0.713 (0.067)	0.510 (0.035)
	SVM	0.894 (0.043)	0.631 (0.024)	0.512 (0.050)	0.914 (0.035)	0.620 (0.034)	0.525 (0.047)
	SEQC	0.931 (0.02)	0.735 (0.072)	0.544 (0.052)	0.929 (0.02)	0.756 (0.082)	0.563 (0.038)

As demonstrated in this study, HetEnc outperformed previously-reported machine learning models overall, achieving a significantly better predictive performance. Three aspects accounted for its performance: (1) By using exact the same dataset (i.e., the same pre-processed data as model input), HetEnc showed significantly better predicting performance than such machine-learning algorithms as support vector machine (SVM), nearest shrunken centroids (NSC) and k-nearest neighbors (KNN). We observed this superior performance from both "head-to-head" comparative analysis and previously-published results. (2) With no restrictions on data pre-processing and modeling strategies, HetEnc still performed better than the best models developed by other groups in the SEQC project. (3) Performance differences between cross-validation and external testing are relatively small in developed HetEnc models, indicating that the HetEnc model can be applied more generally to new test sets.