Prediction of protein isoforms using semi-supervised learning

Álvaro Gutiérrez León s212714, Baris Kara s213449 and Casper Rasmussen s206220
Course: 02456 Deep Learning, Project 29
All students contributed equally to this work

Background & Aim

RNA-seq can be used to snapshot the transcriptome of cells and the ample data that is being generated with this technique. However, the entire cellular expressome includes all the protein isoforms being produced by the cell, which RNA-seq does not elucidate. Hence, this project investigates whether the protein isoform expression profile can be predicted from gene expression profiles using a semi-supervised machine learning approach. First, a Variational Autoencoder (VAE) was trained on large amounts of high-dimensional RNA-seq data to obtain a model that is able to provide informative low-dimensional encodings of the gene expression profiles. A smaller gene expression dataset, accompanied by a protein isoform expression dataset, was then encoded with the trained VAE model, and the encoded variables were fed into a feed-forward neural network in order to predict protein isoform expression.

Data specifications

The ARCHS4 and GTEX gene expression datasets were log2-transformed with a pseudo-count of 1 and quantile normalized, and then modelled using a hurdle model:

$$prob(x = 0) = p$$
$$prob(x > 0) = (1 - p) \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

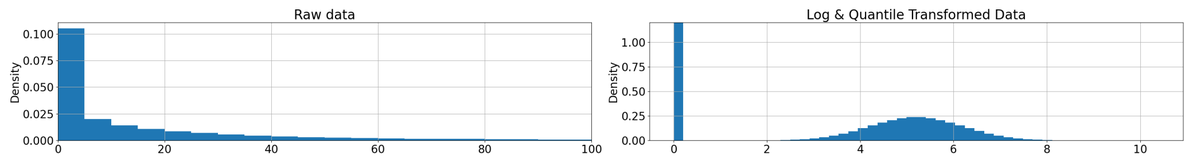

Figure 1: Transformation of raw data into hurdle distribution.

Models

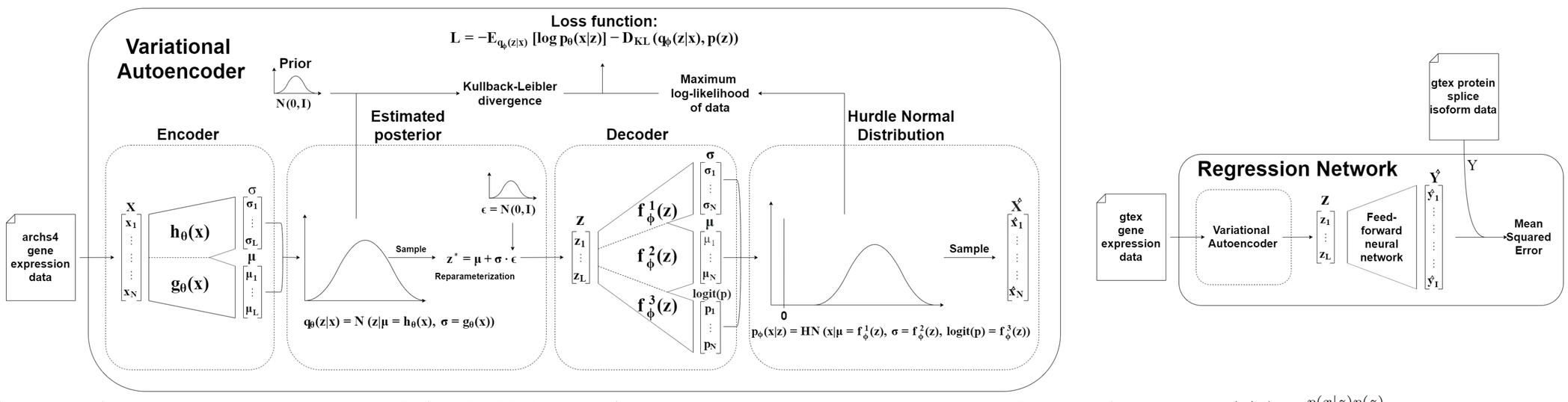

Figure 2: Overview of the data pipeline and models. Left: The Variational Autoencoder can be understood through a Bayesian framework, $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, where the encoder estimates the posterior $q_\theta(z|x)$. We assume a prior distribution $p(z)$ which follows a standard normal distribution $N(0, I)$ in order to regularize our latent space. The decoder parameterizes the observation model which returns the likelihood $p_\phi(x|z)$, that is, the likelihood of the data. Right: The regression network takes gene expression data encoded into latent dimensionality and predicts protein isoform expression through a feed-forward Neural Network.

Training the model

The VAE was trained on the ARCHS4 dataset, which featured approximately 170 thousand samples and 35 thousand genes. Grid search was used to optimize the hyperparameters and, consequently, the representational power of our latent space encodings. During training, the Evidence Lower Bound (ELBO), Kullback-Leibler Divergence and the log likelihood of the data were monitored as shown in Figure 3 (top). Our final ELBO training and validation loss were -8720 and -9271, respectively. To visualise the reconstruction of our 34558-dimensional data we generated a heatmap of both the initial input $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$ as shown in Figure 3 (middle).

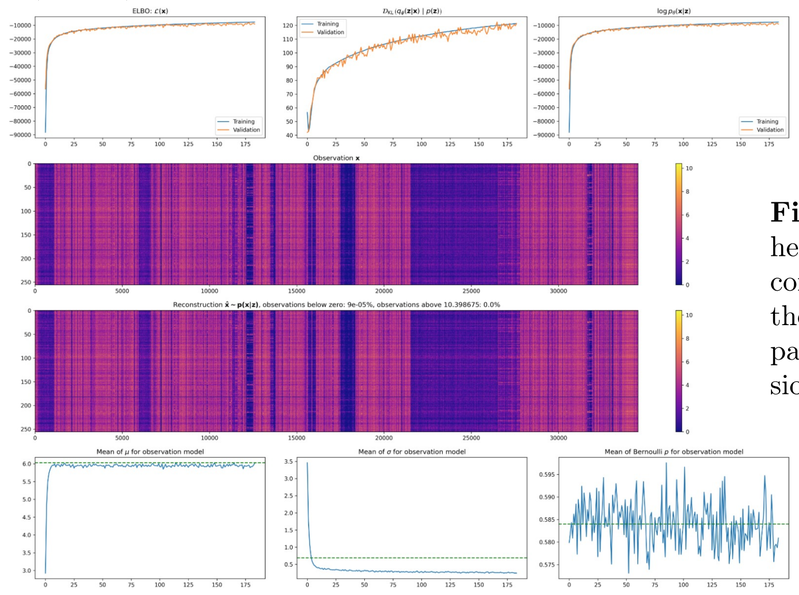

Figure 3: Loss curves, heatmap visualization of reconstruction and tracking of the mean of observation model parameters over all dimensions.

The ability of the VAE to capture the data distribution of the input data through the observation model was also monitored, by plotting the mean of the parameters of the observation model and comparing to the same parameters estimated directly from the input dataset. As can be seen from Figure 3 (bottom), the model obtains a fairly good estimate of the $logit(\mathbf{p})$-parameter and the weighted mean μ of the data, but undershoots the standard deviation σ to some extent. The reconstructed data was also evaluated on its cosine similarity to the input, and it resulted in a value of 0.88.

Latent space evaluation

The GTEX gene expression data, consisting of roughly 17 thousand samples and 34558 genes, was input to the trained VAE model and a latent vector with 64 dimensions was obtained. UMAP dimensionality reduction was applied to both the original data and the VAE encodings. From this it appears that the VAE encoding has slightly better separation of tissue classes. As such VAE encoding appears to not only preserve most of the original information but also orders it in a way that makes distinct clusters more separable.

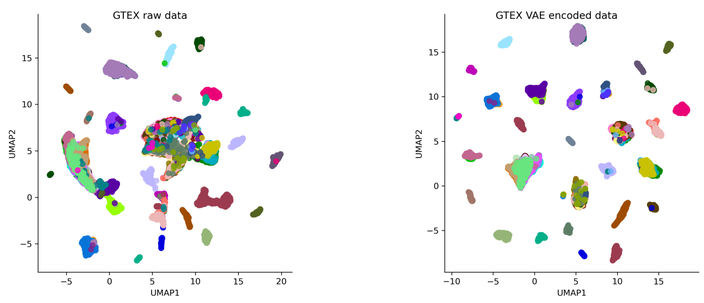

Figure 4: UMAP dimensionality reduction of original data and of VAE encodings.

Protein splice isoform predictions

As a naïve baseline for making predictions about isoform expression based on gene expression, we performed Lasso regression (alpha=1) resulting in a poor fit with $R^2 = 0.091$. The gene expression encodings produced in the earlier step were then used as input in a regressor model comprised of a simple feed-forward neural network. A dataset containing the expression profiles of 965 different isoforms in the 17 thousand samples was used as label for the regressor. Additionally, as a baseline model, a principal component analysis (PCA) of the GTEX gene expression data was conducted, and its components were also used as input to the regressor.

The loss was measured by mean-squared error (MSE) of the predicted isoform expression vector versus the label vector, and it rapidly converged to values close to 0 (Figure 5, left). The datapoints of the prediction and the target were plotted against each other, and the Pearson correlation was calculated, showing excellent values. The results showed a similar performance of the regressor model based on VAE encodings and the PCA-based regressor (Figure 5, right).

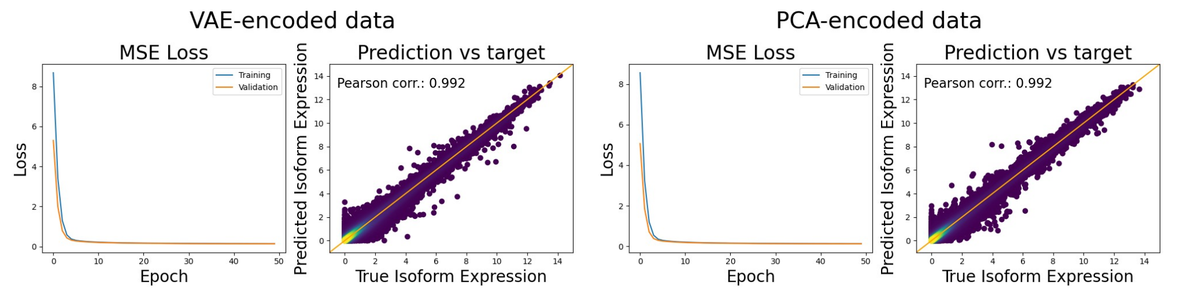

Figure 5: Training curves and prediction plots for both VAE and PCA encoded data.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pearson corr.	0.993	0.989	0.991	0.992	0.987	0.991	0.985	0.994	0.990	0.992	0.993	0.990	0.981	0.991	0.990	0.992

Table 1: Per sample Pearson correlation coefficient between real and predicted 965-dimensional isoform expressions for a batch of 16 VAE-encoded test samples.

Discussion & Conclusion

This project illustrates that VAEs can be used in a semi-supervised machine learning pipeline by taking a larger dataset with no labels to train on, and then using the trained model to generate informative representations of the underlying data structure in a smaller dataset. When these smaller datasets have target labels, supervised learning can be performed. Thus, the entire model learns from the structure of larger and simpler datasets to enable powerful predictions on smaller but more extensive datasets. In the present case the encodings from the VAE did not outperform a simple PCA approach. It could be that the complexity of the data is not high enough to warrant using the VAE approach, or it could be due to wrongful assumptions about the data during modelling. As our model barely overfits it could also be that with more time for tuning it would eventually surpass simpler approaches.

References

1. Lachmann, A., Torre, D., Keenan, A.B. et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun 9*, 1366 (2018). https://doi.org/10.1038/s41467-018-03751-6

2. Lonsdale, J., Thomas, J., Salvatore, M. et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet 45*, 580–585 (2013). https://doi.org/10.1038/ng.2653