LDEncoder: Reference deep learning-based feature detector for transfer learning in the field of epigenomics

Gun Woo (Warren) Park*
Dept. of Computer Science, University College London
London, UK
warren.park.17@gmail.com

Kevin Bryson
School of Computing Science, University of Glasgow
Glasgow, UK
kevin.bryson@glasgow.ac.uk

ABSTRACT

We propose a reference feature extractor that can be used for methylation data and potentially other epigenomic data sources. In doing so, it can be used in a trans-omics manner to bridge between epigenomics and transcriptomics. By having an internal latent space, it can solve classification/regression problems in a trans-omics manner. DNA methylation data is part of epigenomics data that is altered by external factors including the change in environment. It has multiple roles including the regulation of gene expression. The goal of the reference feature extractor is to extract important features from the DNA methylation data while encoding the features in a low dimensional feature space. To achieve this, a pan-cancer dataset was used to train the model with a wide variety of data. Due to the low dimensional encoding, downstream tasks can be solved while utilising significantly fewer parameters. The current state-of-the-art can work with a trans-omics setting, but it was not able to generalise the model so that it could work in other settings [1–3]. For example, TDImpute [4] needed an extra decision-making model to complete the classification task, while not utilising the latent feature representation inferred inside the model. Furthermore, a multi-layer perceptron, called LDEncoder, used in this approach has a low encoding dimension (512), which is used to represent the high dimensional DNA methylation data in a significantly lower-dimensional feature space. So, if the new classification/regression problem needs to be solved, the input dimension of 512 can be used for the transfer learning of the model. This significantly reduces the amount of time and computational resources needed for solving problems. In effect, transforming the DNA methylation data to gene expression data (RNA-seq) while having a bottleneck enables the lower dimensional encoding of the data. Also, in a similar scenario, we evaluated the performance of various models and techniques inspired by successful ones in computer vision. These included incorporating the model parameter savers based on the best validation loss and CpG site sorting[1]. We found some promising results as shown in Table 1. Also, we further evaluate the generalisability of the model through cancer/non-cancer prediction and breast cancer molecular subtype prediction results.

[1]This allows the locality features on the data to be exploited

BCB '21, August 1–4, 2021, Gainesville, FL, USA

ACM ISBN 978-1-4503-8450-6/21/08.
https://doi.org/10.1145/3459930.3469487

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**.

KEYWORDS

epigenomics, transcriptomics, transfer learning

ACM Reference Format:
Gun Woo (Warren) Park and Kevin Bryson. 2021. LDEncoder: Reference deep learning-based feature detector for transfer learning in the field of epigenomics. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3459930.3469487

	RMSE	R^2
LDEncoder 600[2]	0.6496± 0.0048	**0.9414± 0.0009**
ResNetSorted[3]	0.6568±0.1240	**0.9402±0.0022**
Baseline	0.6566±0.0030	**0.9400±0.0007**
ResNet	0.6608±0.0085	**0.9395±0.0014**
LDEncoder	0.6725±0.0050	**0.9371±0.0008**
ResNetCHR[4]	0.6778±0.0083	**0.9362±0.0014**
VGGSorted[5]	0.7549±0.0054	**0.9217±0.0011**
VGG	0.7696±0.0071	**0.9186±0.0015**
Baseline 4096[6]	0.9201±0.0117	**0.8828±0.0031**
PCA	4.1664±0.0248	**0.2750±0.0047**

Table 1: Results for each candidate reference model. The ranking given is based on the squared Pearson's correlations coefficient.

ACKNOWLEDGMENTS

We thank TCGA Research Network for providing the necessary datasets required for the completion of this research work.

REFERENCES

[1] Karim Malki, E Koritskaya, F Harris, K Bryson, M Herbster, and MG Tosto. Epigenetic differences in monozygotic twins discordant for major depressive disorder. *Translational psychiatry*, 6(6):e839–e839, 2016.
[2] Xuesi Dong, Lijuan Lin, Ruyang Zhang, Yang Zhao, David C. Christiani, Yongyue Wei, and Feng Chen. ToBMI: Trans-omics block missing data imputation using a k-nearest neighbor weighted approach. *Bioinformatics*, 35(8):1278–1283, 2019.
[3] Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A. Salas, and Brock C. Christensen. MethylNet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics*, 21(1):1–15, 2020.
[4] Xiang Zhou, Hua Chai, Huiying Zhao, Ching Hsing Luo, and Yuedong Yang. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning-based neural network. *GigaScience*, 9(7):1–10, 2020.

[2]LDEncoder model with 600 epochs
[3]ResNet inspired model with sorted dataset
[4]ResNet inspired model with sorted dataset with chromosomal separation of the data
[5]VGG inspired model with sorted dataset
[6]Baseline experiment with batch size of 4096