Uncovering the gene usage of human tissue cells with joint factorized embeddings

Assya Trofimov [1 2] **Joseph Paul Cohen** [1] **Claude Perreault** [3 2] **Yoshua Bengio** [1]
Sebastien Lemieux [1 2]

Abstract

We present a factorized embedding method, that learns simultaneously gene and sample embeddings in their respective latent space, in a gene function-dependent manner. Running the model on RNA-Seq data, we observed that tissue samples aggregated spontaneously in latent space by tissue similarity while genes aggregated in latent space by gene function. Our method recovered most gene-gene association reported in reference databases, but also recovered gene associations found in the literature but absent in databases. Our approach has the potential to uncover not yet known gene interactions, in a tissue-specific manner.

1. Introduction

We propose a deep learning-based dimensionality reduction method that simultaneously learns about samples and genes, while embedding them into their respective latent space based on function. While RNA-seq is a powerful tool to gain insight into a cell's mechanisms, it has many pitfalls, most of which are mostly based on the size of the data. Indeed, an RNA-Seq yields for each studied sample a long vector of gene expression values $(10^4 - 10^5)$. Often analysis is done only on samples to monitor gene expression changes using some type of selection scheme, and to those selected genes are associated functions, through enrichment analysis pipelines. While these methods seem to perform well, they are subjected to stringent cutoffs and are often not reproducible in other datasets (Boutros et al., 2009). Moreover, rare are the methods that examine the gene changes for all samples.

Our contributions:

- We present a method that aggregates tissue samples by similarity in latent space.
- Our method also simultaneously aggregates genes based on their function.
- Our gene embedding space represents gene function and recovers most of reported co-expressions in *ts-coexp* and *GOterms* as well as additional relationships.

[1]Montreal Institute for Learning Algorithms (MILA), Universite de Montréal [2]Institute for Research in Immunology and Cancer, Montreal, Qc [3]Department of Medecine, Universite de Montréal. Correspondence to: Assya Trofimov <assya.trofimov@umontreal.ca>.

2. Factorized Embedding

The core of our approach relies on finding good embedding coordinates for gene and tissue samples in order to minimize the prediction error of the network. We create two spaces where gene and tissue samples are represented. The network receives as input pairs of indices; for a gene expression matrix with N samples by M genes, the network will receive $N \times M$ pairs of indexes and attempt to predict the gene expression for each index pair. The coordinates in gene space and sample space for each of these indices are concatenated and fed through one fully-connected non-linear layer of the neural network. The output is a single neuron, predicting the gene expression value for the input pair and the network is penalized on a Mean Squared Error (MSE) (Figure 1). To contrast our method with bi-clustering (Madeira & Oliveira, 2004), we factor by concepts we control, namely genes and tissue types. Moreover, we found that the embedding space captures functional relationships between genes as well as tissues, something not guaranteed in bi-clustering.

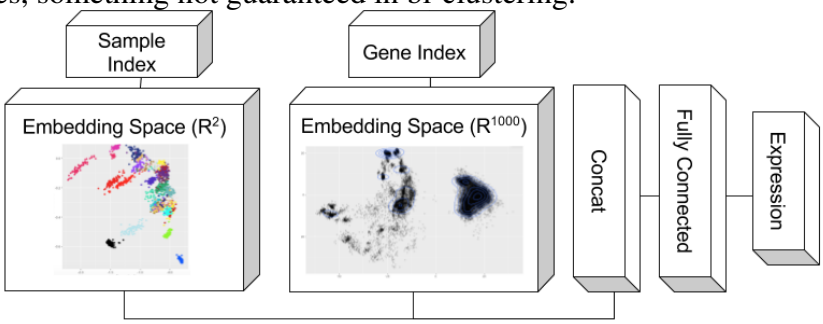

Figure 1. The network architecture

We kept the network above the embeddings minimal, in order to maximize the encoding within the embeddings, rather than the parameters of the fully connected layer. We have tested increasing sizes of gene and sample embedding space, but only changing gene embedding space seemed to significantly improve performance. For all experiments, the fully connected layer contains 10 or 25 neurons. Training is done by iterating through every pair of sample and gene indexes and adjusting the embeddings and network parameters to decrease prediction error. We train using a batch size of 10,000 over the entire dataset of 112M pairs, with an RMSprop optimizer, at a learning rate of 10^{-3}. Our code was implemented using Theano (Theano Development Team, 2016) and Keras (Chollet & Others, 2015) and we release our source code online[1].

3. Experiments

3.1. Data

We use the *GTEx* daataset, obtained from the Genetype-Tissue expression project (GTEx), a RNA-Seq dataset, that contains 8910 tissue samples, over 30 different tissue types (Lonsdale et al., 2013). We removed two tissue types, *Brain* and *Fallopian Tube*, mainly for the sample size and ambiguous labeling. Gene expression values are continuous values, that have been log-transformed for this analysis.

3.2. Tissue embeddings

We first examined the tissue embedding space. We found that our method as well as t-SNE separate the samples into groups (Figure 2). We then investigated how these tissue groups relate to each other

[1]Source code not released yet.

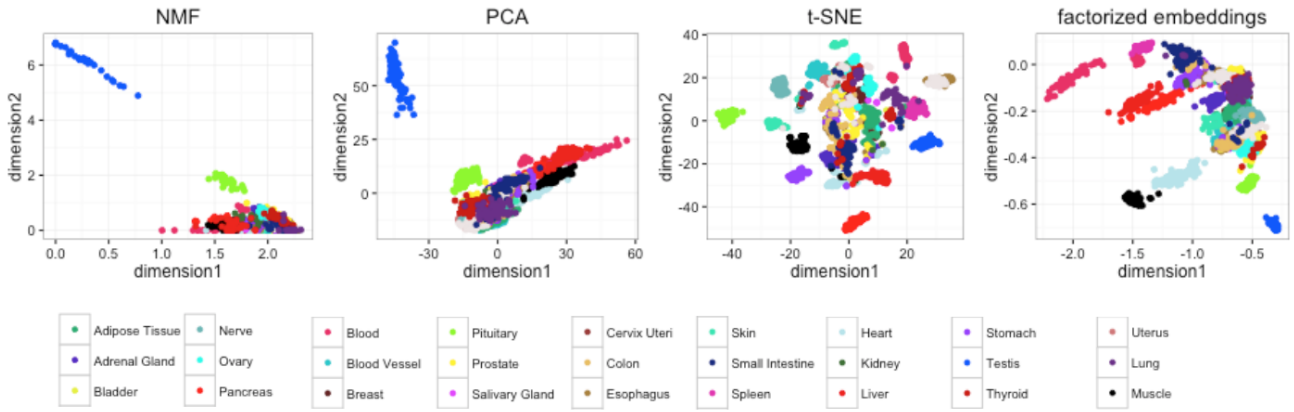

Figure 2. Embedding comparison with popular dimensionality reduction techniques

Algorithms	NMF	PCA	t-SNE	Factorized Embeddings
1-NN	0.31± 0.02	0.38± 0.03	0.62± 0.04	0.62± 0.02
Decision Tree	0.28± 0.03	0.32± 0.04	0.50± 0.05	0.50± 0.05
Gaussian NB	0.22± 0.04	0.22± 0.04	0.28± 0.04	0.29± 0.05
Random Forest	0.28± 0.03	0.31± 0.03	0.47± 0.04	0.47± 0.04
AdaBoost	0.16± 0.04	0.18± 0.04	0.25± 0.06	0.27± 0.06
Vanilla MLP	0.15± 0.02	0.33± 0.02	0.48± 0.03	0.13± 0.02

Table 1. Semi-supervised classification of tissue source for different methods on embeddings. We take these embeddings (which are trained on 2000 samples) and train a classifier using only 2% and present the accuracy of the classification on the remaining 98%. The numbers shown are the average result of 100 random splits with the exception of raw data which was only computed over 10 random splits due to complexity.

in the sample embedding space. Four genes or gene groups were chosen as representative examples; i) keratin genes (KRT) for different epithelial tissues ii) MYH6 for muscle tissues (heart and muscle), iii) CD8B as a marker for immune cells and iv) XIST as a sex-determining gene (Figure 3). Our results suggest that the factorized embeddings, but not t-SNE, segregate space into regions of tissue similarities, according to gene expression values. We also note that expression of sex-related genes is not a major contributor to localization in embedding space.

We benchmarked our method against other dimensionality reduction techniques, such as principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten, 2014) and non-negative matrix factorization (NMF). We observe that t-SNE (at perplexity 30, chosen via a hyper-parameter search) and our method generate embeddings that contain similar intra-group information (Table 1).

3.3. Distance vs distribution of linked genes

To account for the observed better division of tissue embedding space by our method, we hypothesized that the performance is due to the model simultaneously learning about gene functions in cells. Similar to what was observed in tissue embedding space, we expected the model to aggregate genes participating in similar functions in the gene embedding space.

To verify this, we have compared the euclidean distance between gene embeddings to two gene function databases: i) *Gene Ontology* (Ashburner et al.), a gene function annotation database, manually curated, assembled as a hierarchical graph, and, ii) *ts-coexp*, a gene co-expression graph, based on the top 1% correlated genes over human tissues as well as cancer cell lines (Piro et al., 2011). We also ran the *ts-coexp* pipeline described in (Piro et al., 2011) on our own dataset, to account for the

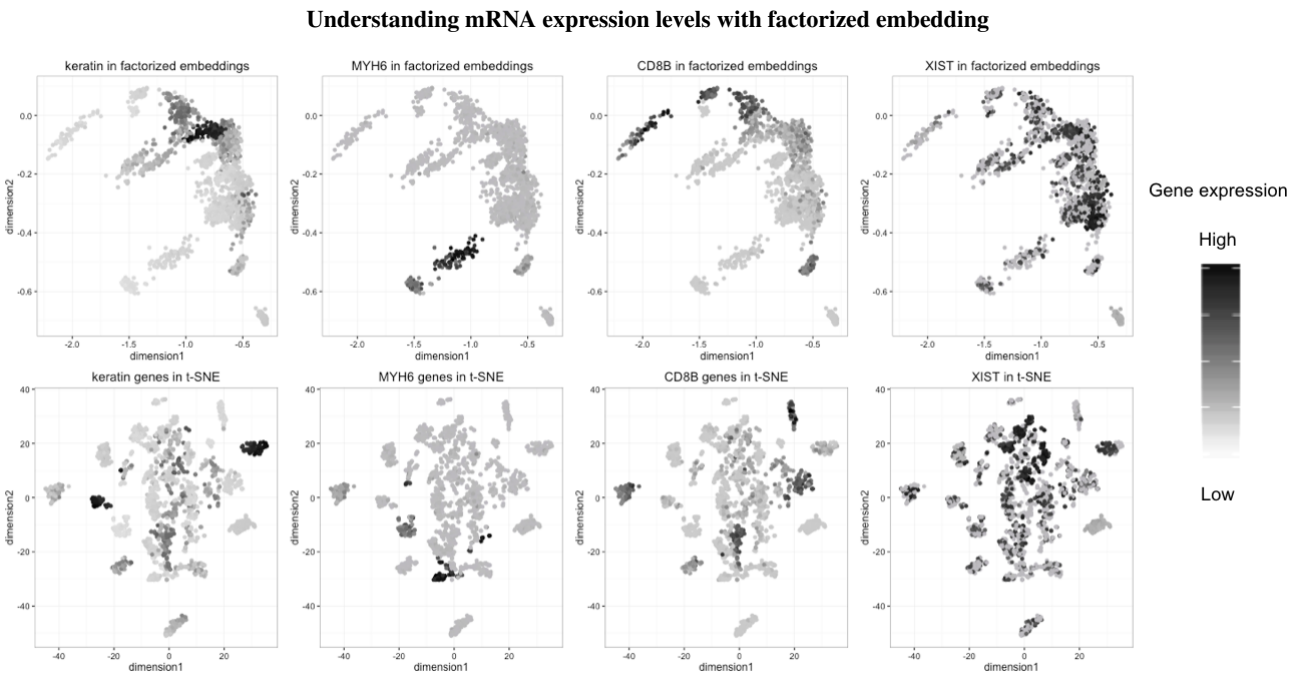

Figure 3. Embedding plots coloured by relative gene expression value. Factorized embeddings but not t-SNE exploits tissue embedding space as a function of relative gene expression, as seen by a gradient of darker colors across different tissue clusters.

minor dataset difference (*ts-coexp* authors included cancer cell lines in their analysis, whereas we only kept to healthy human tissues). We report that, as expected, proximity in gene latent space coincides with participation in similar functions (genes found in the same *GO terms*) as well as general tissue co-expression (reported correlation edges in *ts-coexp*) (Figure 4).

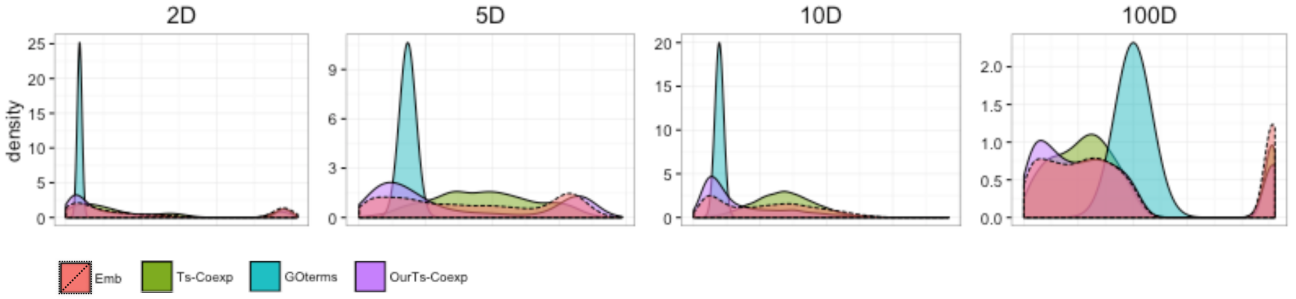

Figure 4. Gene embedding space is organized such that close together genes share similar functions. Here, we show overlap between relationships found in gene space and reference databases, at various gene embedding space sizes. *Emb = factorized embeddings; Our Ts-Coexp = ts-coexp algorithm ran on our dataset*

Finally, we overviewed the discrepancies, where predicted close together genes were absent from the reference databases. Most pairs are pseudogenes, snoRNA and miRNA, which, although not properly studied in the literature, seem to correlate in our dataset.

4. Conclusion

In this work we have shown that our method learns simultaneously i) sample embeddings by stratifying the space by tissue similarity and ii) gene embeddings, by aggregating genes participating in common functions. Finally, because of the way the data is input, our method is robust to partial data. This allows for incorporation of many datasets or single-cell RNA-Seq experiments.

References

Ashburner, M, Ball, C A, and Blake, et. al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*. ISSN 1061-4036.

Boutros, Paul C, Lau, Suzanne K, and Pintilie, Melania et. al. Prognostic gene signatures for non-small-cell lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 2009. ISSN 1091-6490.

Chollet, François and Others. Keras. https://github.com/fchollet/keras, 2015.

Lonsdale, John, Thomas, Jeffrey, and Salvatore, Mike et. al. The Genotype-Tissue Expression (GTEx) project. 2013.

Madeira, Sara C. and Oliveira, Arlindo L. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2004. ISSN 1545-5963.

Piro, Rosario Michael, Ala, Ugo, and Molineris, Ivan et. al. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. *European journal of human genetics : EJHG*, 2011. ISSN 1476-5438.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, 2016.

Van Der Maaten, Laurens. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 2014. ISSN 1532-4435.