Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders

Way GP and Greene CS

Poster: Jordi Esteve and Stathis Fotiadis

Introduction

- The goal is to learn a model that enables the identification of biologically relevant patterns from data.
- This is done by mapping data to a latent space using a Variational Autoencoder (VAE).
- Patterns in the latent space be used to predict tumor's response to therapy or characterize the gene expression in different tumors.
- Data are RNA-seq gene expressions (transcriptomes) extracted from The Cancer Genome Atlas [1]. In total ∼10k tumors from 33 different types of cancer.
- Transcriptome data were selected because it is established that can describe tumor states.

Model

- The model is a VAE [2] named Tybalt that encodes the 5,000 most variable expressed genes to a 100 dimensional subspace.
- Advantages of VAEs is that it is unsupervised and represent non-linear relationships stochastically.
- Biological data have noisy labels so VAE is a good fit for the task.

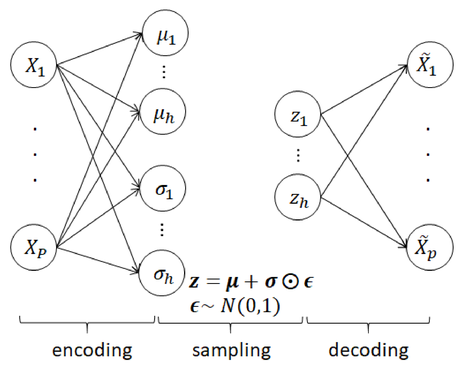

Figure 1: Architecture of a single layer VAE. Figure adapted from [3].

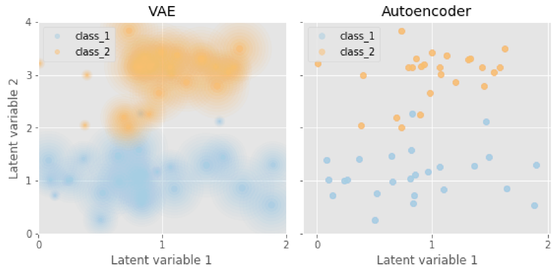

Figure 2: Comparison of VAEs and Autoencoders. In VAEs the latent space is designed to be stochastic which helps with noisy data.

Conservation of information

- t-SNE [4] is used to reduce both the original data and the latent features down to two dimensions.
- As seen in Figure 3, the clusters of different cancer types and the relations between them are maintained between the original and latent features.

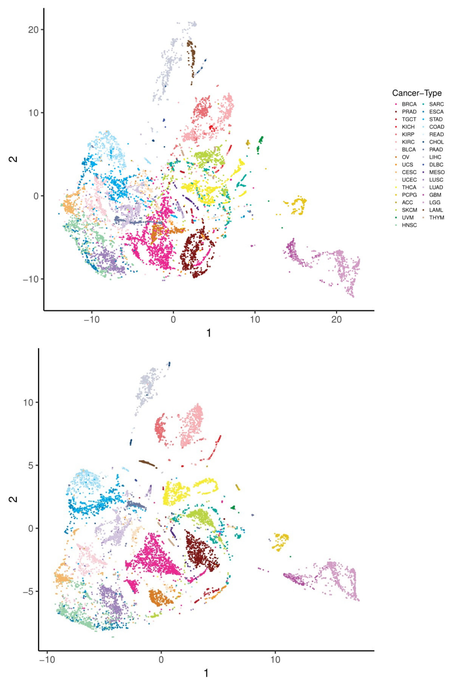

Figure 3: Top: t-SNE of latent information. Down: t-SNE of original data. Figure from [3].

Biological signal in encodings

- The latent space encapsulated biological signal.
- Patient gender was almost perfectly distinguished by encoding 82 (Figure 4).

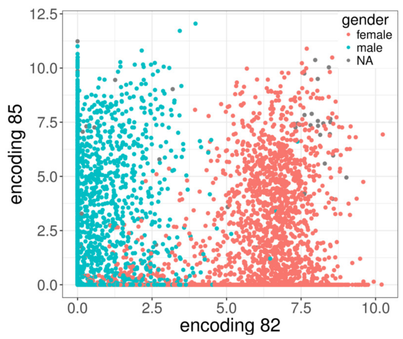

Figure 4: Encoding 82 stratified patient sex. Figure extracted from [3].

- Genes related to gender like ones coming from Y chromosome were found to have decoder weight with high absolute values.
- Encodings 53 and 66 were able to distinguish between skin cutaneous melanoma tumor (SKCM) types (Figure 5).

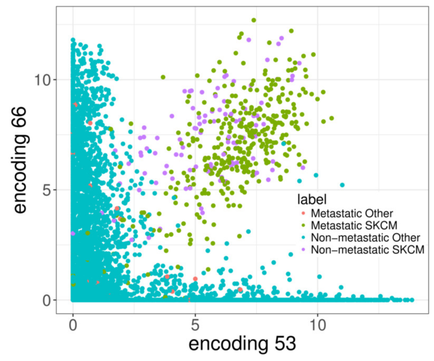

Figure 5: Encodings 53 and 66 separated SKCM melanoma types. Figure extracted from [3].

Latent space interpolation

- The difference between related subtypes of cancer was demontstrated in the case of high grade serous ovarian cancer (HGSC).
- Mesenchymal and immunoreactive subtypes can be distinguished by gene expression.
- Encodings 87 and 56 showed the largest differences between the two subtypes (Figure 6 A&B).
- Encoding 87 is associated with collagen generation and extracelular matrix organization processes known to be related to the mesenchymal subtype (Figure 7).
- Encoding 56 is related to immune system responses and the immunoreactive subtype (Figure 7).

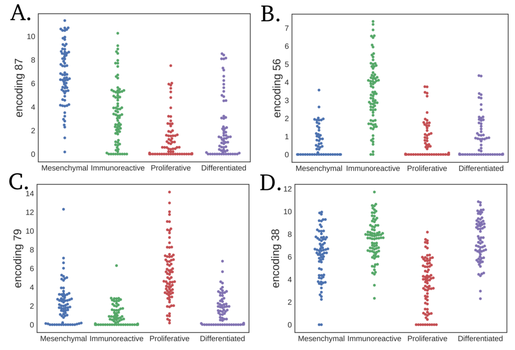

Figure 6: Encodings with the maximum difference between the mesenchymal/immunoreactive(A&B) and the differentiated/proliferative(C&D) HGSC subtypes. Figure extracted from [3].

Encoding	Tail	Subtype Enrichment	Pathway
87	+	Mesenchymal	Collagen Catabolic Process
87	+	Mesenchymal	Extracellular Matrix Organization
87	-	Immunoreactive	Urate Metabolic Process
56	+	Immunoreactive	Immune Response
56	+	Immunoreactive	Defense Response
56	+	Immunoreactive	Regulation of Immune System Process
56	-	Mesenchymal	*No significant pathways identified*
79	+	Proliferative	Chemical Synaptic Transmission
79	-	Differentiated	Xenobiotic Glucuronidation
38	+	Differentiated	*No significant pathways identified*
38	-	Proliferative	Xenobiotic Glucuronidation

Figure 7: Association encodings to known biological pathways in HGSC. Extracted from [3].

- Analogously, the differentiated and proliferative tumours are distinguished by the encodings 38 and 79 (Figure 6 C&D).
- Both encodings are related with glucuronidation, a process connected to the response to chemotherapy and survival rate in colon cancer.
- Encoding associations for known biological pathways per HGSC subtype can be seen in Figure 7.

Conclusions

- Tybalt has shown promising results in capturing biologically relevant encodings.
- Decoder weights were found to encapsulate gene ontology of patient gender and could reveal further dichotomies.
- Encodings were linked with already indentified biological pathways in HGSC subtypes.
- Further careful validation and thorough evaluation is needed. Unexplained encodings should be investigated for unknown pathways.
- Generalization of the methodology in less heterogenous datasets could also prove beneficial.

References

[1] K. Tomczak, P. Czerwinska, M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge", Contemporary Oncology, 2015

[2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes", arXiv:1312.6114, 2013

[3] G. P. Way and C. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders" Pacific Symposium on Biocomputing, 2018

[4] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, 2008