Project Proposal

IITM-CS6024 : Algorithmic Approaches to Computational Biology

Team name: **BioBois**
Project Title:
DeepVAE: A deep learning approach for latent space modelling of cancer gene expression data.

Members: Rahul Nikam BT18D011 and Anoubhav Agarwaal BE16B002

Introduction/Motivation:

Understanding cancer transcriptomic data is crucial for stratifying prevalent tumor types. If we can extract the hidden underlying representation from the gene expression profile, we can use it to predict a tumor's response to specific targeted therapies. Inferring the hidden distribution of data falls under the family of Unsupervised Learning algorithms. Generative models are one such type of algorithms, which have been used in vision and natural language to generate realistic-human level visuals and text. It can be hugely transformative if it can do the same for biomedicine. Extracting a biologically relevant latent space is the first critical in our endeavor to demystify cancer. Hence, we had picked an influential paper by Way and Greene(2018)[1], for our paper presentation. Our primary research objective is to extend the results of Way and Greene(2018)[1] to obtain a better biologically relevant latent space using Deep Variational Auto-Encoder. The learned representation is evaluated based on the recapitulation of biological signals in the t-SNE plots and pan-cancer classification accuracy. The pan-cancer classification accuracy is compared with state-of-the-art CNN methods. We also perform model interpretability through SHAP analysis and others. The contributing genes and its significant pathways are presented for all tumor types.

Related work:

- Way and Greene(2018)[1] uses a shallow VAE for extracting biologically relevant latent space from cancer gene expression data. The decoder weights of the shallow network were directly used to identify the contributions of significant genes. Gene Ontology Enrichment analysis was performed to determine significant pathways.
- The authors of [2, 3] mention that they extend Tybalt to learn a latent space for methylation data of lung cancer and breast cancer, respectively.
- [4] Used denoising AE to identify gene signatures related to asthma severity. Enrichment pathway analysis was performed on high-contributing genes. These genes were used to develop supervised random forest classifier to determine asthma severity.

- [5] Uses VAE on pan-cancer methylation data and performs downstream processed such as cell-type deconvolution, pan-cancer classification, and subject age prediction. They perform SHAP analysis to determine the importance of CpGs.
- [6] An influential paper of pan-cancer classification using K-nearest neighbors/Genetic Algorithms. They achieved an accuracy of 90%.
- [7, 8] Recent deep learning-based pan-cancer classification achieved state of the art performance of 96%. They employ Convolutional Neural Networks on the gene expression data.

Contributions:

1. To extract and characterize the latent space obtained using deep VAEs.
2. To extend model interpretability using SHAP analysis and others.
3. To develop a deep VAE framework for any disease gene expression dataset.

The **methodology** is covered in the expected outcomes section.

Expected result/outcomes:

1. The successful extension of Tybalt to deep variational autoencoders.
2. To gauge the usefulness of deep VAE compression at different latent vector dimensions.
3. To produce higher quality biologically relevant representations (or latent space) than Tybalt.
4. The t-SNE plot should retain more biological signal than Tybalt,
5. Higher pan-cancer classification accuracy over 33 prevalent tumor types.
6. To employ transfer learning on the encoder network for the downstream prediction tasks.
7. To classify patient sex using the encoder network. To identify and confirm that the contributing genes correspond to those located on the sex chromosome.
8. To perform a comparison (accuracy) between the state-of-the-art pan-cancer classifiers (based on convolutional neural networks) and the deep encoder network.
9. To extend model interpretability for deep VAE based on SHAP and LIME analysis and identify the contributing genes for each tumor type.
10. To report the significant pathways using Gene ontology enrichment analysis.

Timeline/workload:

Timeline	Rahul	Anoubhav
4-9 October	Replication and code Deep VAE	
10-11 October	Modification of VAE model to classification using SVM, RF etc.	t-SNE, clustering
12-13 October	-	find significant genes using SHAP or similar library
14-15 October	Enrichment analysis; for the SHAPS.	
16-17 October	Getting Hypothesis to significant genes	
18 October	Buffer	
19-21 October	Research on possible improvement in code	
22 October	Meeting Day to decide further work	
23-27 October	Work on the individual task assigned in meeting for 5 days	
28-29 October	Buffer	
30 October - 3 November	Minor additional experiments and PPT/report making	

References

[1] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," *bioRxiv*, 2017.

[2] Z. Wang and Y. Wang, "Exploring DNA Methylation Data of Lung Cancer Samples with Variational Autoencoders," *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 1286–1289, 2019.

[3] A. J. Titus, O. M. Wilkins, C. A. Bobak, and B. C. Christensen, "Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction," *bioRxiv*, p. 433763, 2018.

[4] S. Lou, T. Li, D. Spakowicz, G. L. Chupp, and M. Gerstein, "Latent-space embedding of expression data identifies gene signatures from sputum samples of asthmatic patients," 2019.

[5] J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas, and B. C. Christensen, "MethylNet : A Modular Deep Learning Approach to Methylation Prediction," *bioRxiv*, 2019.

[6] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li, "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data," *BMC Genomics*, vol. 18, no. 1, pp. 1–13, 2017.

[7] J. M. D. Guia, "DeepGx : Deep Learning Using Gene Expression for Cancer Classification," pp. 913–920, 2019.

[8] B. Lyu and A. Haque, "Deep Learning Based Tumor Type Classification Using Gene Expression Data," *ACM-BCB 2018 - Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, no. August, pp. 89–96, 2018.