Auto-Encoding for predicting the molecular pathways associated with lncRNAs in cancer

Master's Project Proposal

Molecular pathways determine a cell's every action; including whether to divide, apoptosis, or alter the metabolism. When a cell becomes cancerous many pathways are significantly augmented. To identify which therapies to apply or which patient subpopulations are at the highest risk for cancer re-occurrence due to cancer-related long-non coding genes (lncRNAs), one must understand which pathways are altered and to what extent. In this project, we are interested in learning a subset of lncRNAs specific to sgnalling pathways that are altered between tumor and normal samples. The subset of lncRNAs are ones which are close (<100KB) to transcription start site (TSS) to the cancer associated SNPs and phenotype associated SNPs.

Similar to [1], we will build a hierarchical model that mimics the inherent hierarchical structure of signaling pathways. In contrast to [1], our training data will be gene expression estimates from nearly 13,000 samples from openly available resources such as The Cancer Genome Project (TCGA) and Genotype-Tissue Expression (GTEx) of those subset lncRNAs and protein coding (PC) genes mapped to cancer associated and phenotype associated SNPS. Furthermore, instead of predicting a specific phenotype as done in [1], we will reconstruct the original signal through the use of an autoencoder where the structure is defined by known biological networks. For example, biological network structures can be engineered into the model by grouping genes with known interactions in the first layers of the encoder, similar to [2]. From the trained model, we will examine the latent representation of the lncRNAs to identify if our latent representation reflects known clinical markers. For example, if a set of lncRNAs shows a high amount of KRAS pathway activation it should be correlative to the PC genes associated genetic aberrations in the KRAS pathway. Additionally, we will apply our trained model to several external data sources such as the Cancer Cell Line Encyclopedia and the L1000 dataset, to identify if our latent representation is predictive of therapeutic responses.

In conclusion, we will build a pathway-specific latent representation of cancer associated lncRNAs and validate its utility through external clinical markers and therapeutic responses.

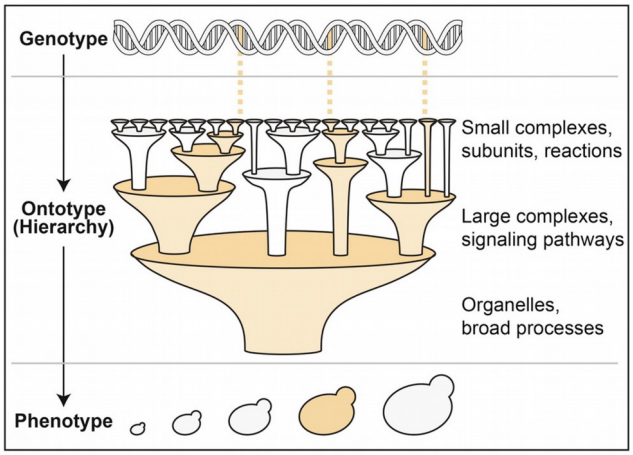

Figure from [1]

Project goals:

- Create a bio-inspired autoencoder
- Identify the pathways specific to the lncRNAs
- Identify if the resulting latent representation is clinically relevant

Prerequisites:

1. Highly experienced in Python programming.
2. Some experience working with TensorFlow.
3. Experience working in a high performance computing system.

Work Plan:

1. Get familiar with cancer pathways and previous approaches
    a. Identify key signalling pathways to be included in the study.
    b. Prepare TCGA and GTEx data.
        i. Additionally CCLE, L1000 if time permits
2. Generate baselines and compare on well defined pathways methods
    a. Create baseline autoencoder without biologically informed structure
3. Engineer novel approach to problem
    a. Create autoencoder with “bag of genes model”, based on the correlation between PC and lncRNAs
    b. Evaluate performance on toy data
4. Apply to cancer data
    a. Correlate pathway activation score with clinical markers
        i. Key pathways of interest: Oncogeneic signalling(HIPPO, PI3K, WNT, RAS, NOTCH,etc), progeny (JAK-stat, MAPK, P53, NFKB, etc.) and metabolic pathways

Contact:

Dr.Alva Rani James
email: Alva.Rani@hpi.de

Reference Material:

1. Yu, Michael Ku, et al. "Translation of genotype to phenotype by a hierarchy of cell subsystems." Cell systems 2.2 (2016): 77-88.
2. Mao, Weiguang, et al. “Pathway-Level Information ExtractoR (PLIER) for gene expression data.” Nature Methods volume 16 (2019), 607–610.
3. Way, Gregory P., et al. "Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas." Cell reports 23.1 (2018): 172-180.
4. Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences 102.43 (2005): 15545-15550.
5. Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. "GSVA: gene set variation analysis for microarray and RNA-seq data." BMC bioinformatics 14.1 (2013): 7.
6. Schubert, Michael, et al. "Perturbation-response genes reveal signaling footprints in cancer gene expression." Nature communications 9.1 (2018): 20.
7. Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." Contemporary oncology 19.1A (2015): A68.
8. Vaske, Charles J., et al. "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM." Bioinformatics 26.12 (2010): i237-i245.
9. Sanchez-Vega, Francisco, et al. "Oncogenic signaling pathways in the cancer genome atlas." Cell 173.2 (2018): 321-337.
10. Subramanian, Aravind, et al. "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles." Cell 171.6 (2017): 1437-1452.
11. Sabarinathan, Radhakrishnan, et al. "The whole-genome panorama of cancer drivers." BioRxiv (2017): 190330.