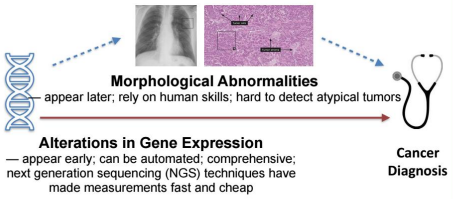
Deep Learning in Pan-Cancer Early Detection

Chao Li and Meng Zhang

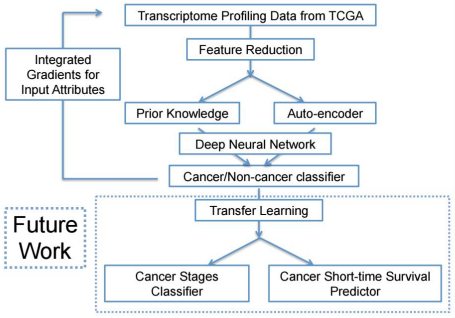
Stanford University

Introduction



Early detection of cancer significantly increases the chance of successful treatment. Current state-of-art cancer diagnosis relies on physicians' experiences to identify morphological abnormalities. Molecular signatures offer an alternative option for early, objective and systematic cancer diagnosis. Deep learning methods are ideal for developing such models as it captures the complicated interactions among different genes.

Procedure



Dataset

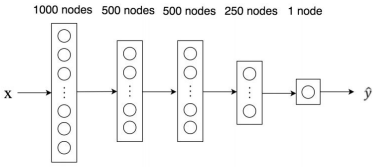
The Cancer Genome Atlas (TCGA) Pan-Cancer Transcriptome Profiling

Sample / Transcript ID	TCGA-OR-A5JX-01	TCGA-HV-A5A5-01	TCGA-MQ-A4LP-01	TCGA-FD-A3N5-01	TCGA-EP-A2KC-01
5S_rRNA	0	0	0	0	0
5_8S_rRNA	0	0	0	0	0
7SK	0	0	0	0	0
A1BG	9.2719	7.2384	9.9624	4.092	16.272
A1BG-AS1	7.9147	6.9539	8.9128	3.1858	6.8324
A1CF	0.9986	8.6651	0	0.6406	13.0142
A2M	14.1457	14.7134	13.1337	11.849	17.3223
A2M-AS1	7.9228	5.628	4.7135	3.4079	6.2567
A2ML1	7.83	2.9618	3.8622	13.6629	0

Illumina HiSeq 2000 RNA sequencing platform, log2 transformed RSEM normalized counts.
58,582 transcripts X 10,663 samples (9,807 cancer and 856 normal), 37 types of cancers

Neural Network

Architecture



Cost Function

$$J = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(a^{(i)}) pos\_weight + (1 - y^{(i)}) \log(1 - a^{(i)})$$

pos_weight to balance precision and recall

Results

Model	Train			Test		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Logistic Regression	7-fold cross validation			0.82	0.80	0.85
SVM				0.79	0.83	0.82
2 Layer NN	0.83	0.95	0.89	0.80	0.89	0.84
4 Layer NN	**0.95**	**0.93**	**0.94**	**0.90**	**0.90**	**0.90**
8 Layer NN	0.75	0.81	0.78	0.80	0.88	0.83

*Hyperparameters: learning_rate = 1e-5, num_epochs = 500, pos_weight = 2

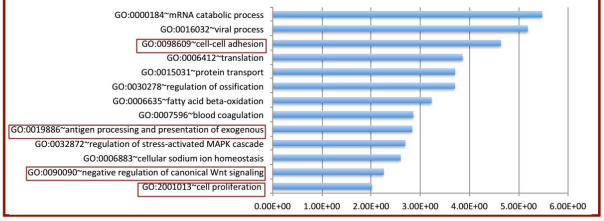
Feature Reduction

	Without Feature Reduction	Prior knowledge (gene sets)				Auto-encoder
		Cell Cycle	Cell Death	Cell Adhesion	Combined	Input / Hidden / Latent Space
Number of Input Features	58,582	1,264	972	1,013	2,907	512
F1 Score	0.90	0.87	0.86	0.85	0.83	-

Interpretation of Neural Network

— Identification of biomarker genes

Integrated Gradients is a method for attributing a neural networks' prediction to its input features by examining the gradients of inputs obtained by interpolating on a straight-line path between the input and a baseline input, and then aggregate these gradients together. The resulting attributions tell us which genes are responsible for predicting cancer.



References

1. Danaee P, Ghaeini R, Hendrix DA. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing.* 2016;22:219-229.
2. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P. Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research 11 (2010) 3371-3408
3. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. arXiv preprint arXiv:1703.01365 (2017).