

Linear classifiers and t-SNE for understanding relationships across cancer types

Ali Yang

October 17th, 2021

Mentors: Alkis Gkotovos and Stefanie Jegelka

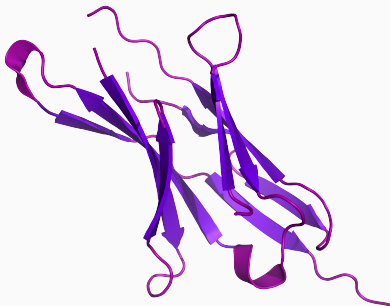
Background

In the US, cancer is the **second** leading cause of death.

How do we treat it?

Background

To treat cancers, we need to be able to classify them.

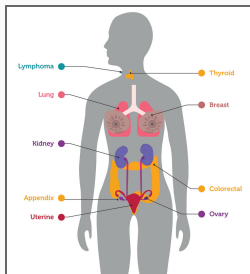


PD-1, a target for immunotherapies¹

¹Image source: Wikipedia

Background

Cancers were originally classified primarily based on the organ or cell they originated from, or growth patterns.



Some of the organs from which cancer can originate²

²Image source: National Cancer Institute

Background

Genetic sequencing has revealed many more cancer subtypes than previously thought.

CMS1 MSI immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high, hypermethylation	SCNA high	Mixed MSI status, SCNA low, CIMP low	SCNA high
<i>BRAF</i> mutations		<i>KRAS</i> mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration, TGF- β activation, angiogenesis
Worse survival after relapse			Worse relapse-free and overall survival

The four consensus molecular subtypes of colorectal cancer³

³Image source: Guinney *et al.*, 2015

How good are existing cancer types?



The Cancer Genome Atlas (TCGA) Dataset

- 9051 patients with known types
- Data on mutation, amplification, and deletion for 763 genes

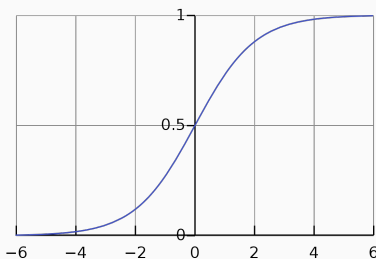
Cancer Type Prediction

- Train a model to predict type from genetics.
- If it has good accuracy, then the types are genetically meaningful.

Cancer Type Prediction

Logistic classifier:

$$f(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}}}.$$

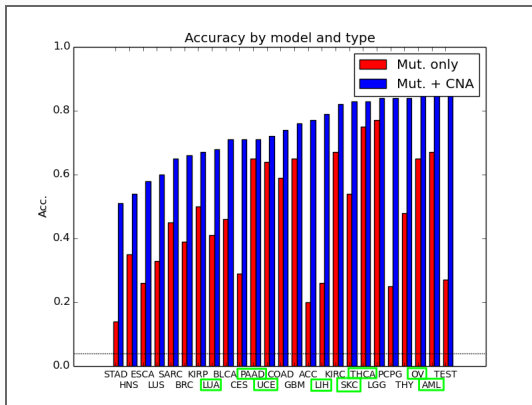


Shape of the logistic classifier's classification boundary⁴

⁴Image source: Wikipedia

Cancer Type Prediction

Overall accuracy: **74.4%**



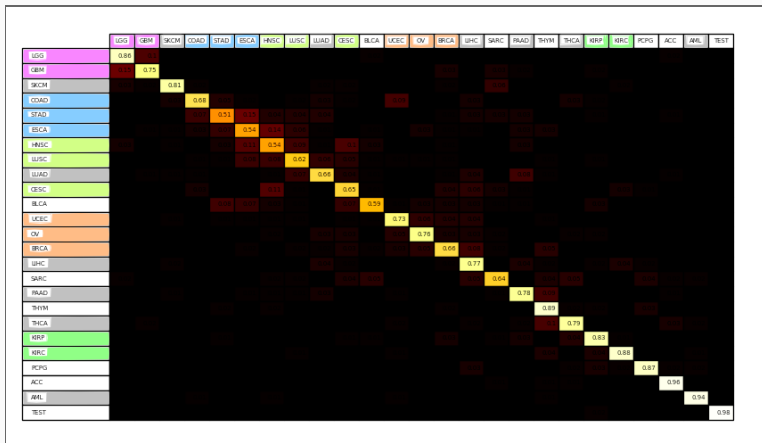
Logistic classifier accuracy on different cancer types

Cancer Type Prediction

	LGG	GBM	SKCM	COAD	STAD	BLCA	HNSC	LUSC	LIAD	CESC	BLCA	LCEC	OV	BRCA	LHCC	SARC	PAAD	THYM	THCA	KIPAN	KIPCN	PCPG	ACC	AML	TEST
LGG	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GBM	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SKCM	0.00	0.00	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COAD	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
STAD	0.00	0.00	0.00	0.00	0.51	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCA	0.00	0.00	0.00	0.00	0.00	0.34	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HNSC	0.00	0.00	0.00	0.00	0.00	0.11	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LUSC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LIAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CESC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LCEC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BRCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LHCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SARC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PAAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
THYM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00
THCA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00
KIPAN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00
KIPCN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00
PCPG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.00
ACC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
AML	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00
TEST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

Logistic classifier confusion matrix

Cancer Type Prediction

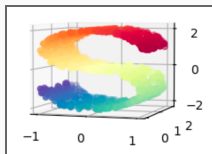


Logistic classifier confusion matrix—related organ groups colored in

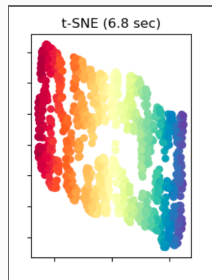
What are other ways to classify cancers?

Classifying Cancers

Dimensionality reduction with t-SNE

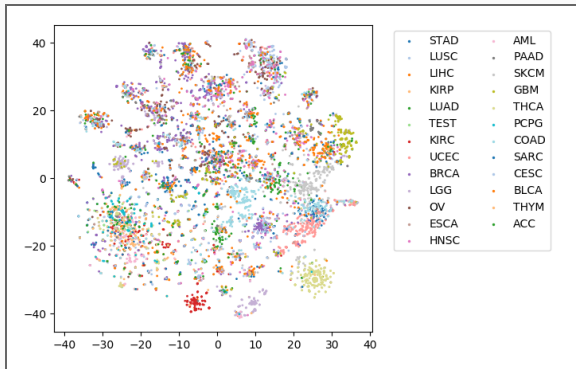


3D data



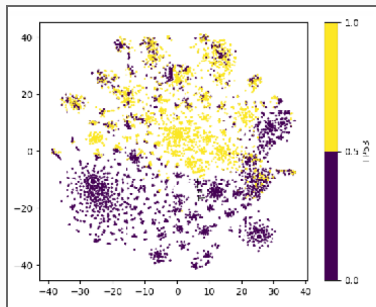
2D version with t-SNE

Classifying Cancers

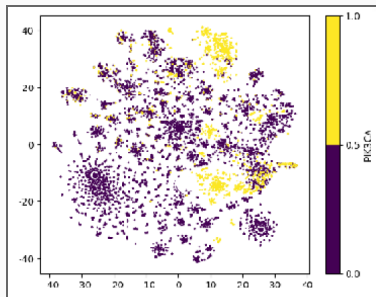


t-SNE of TCGA data, colored by type

Classifying Cancers

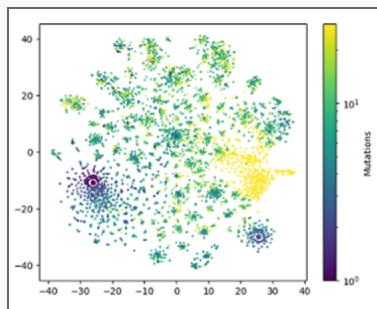


TP53

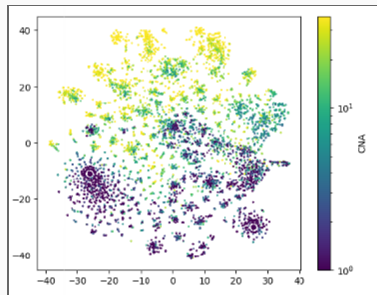


PIK3CA

Classifying Cancers

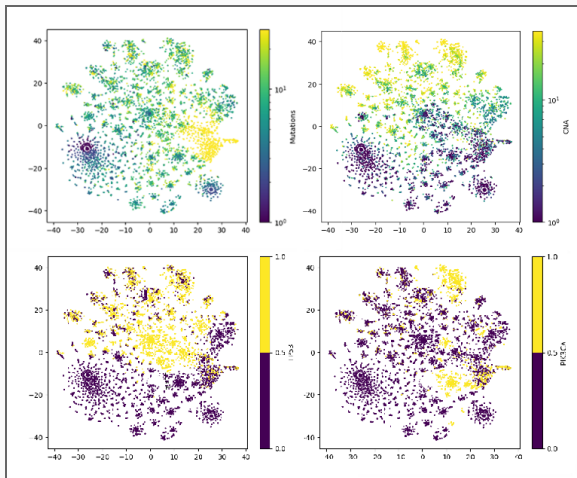


Mutations



Copy-number alterations

Classifying Cancers



t-SNE plot of TCGA data colored four different ways.

From top left clockwise: mutation number, copy-number alteration number, TP53 mutated, PIK3CA mutated

Conclusion

- Existing cancer types can be easily distinguished genetically.
- Cancers from related organs are similar.
- We can classify cancers by TP53 or PIK3CA mutations, number of mutations, and number of copy-number variations.

Acknowledgements

I would like to thank my mentors, Alkis Gkotovos and Professor Stefanie Jegelka, for their guidance throughout this project.

I would also like to thank Qinghong Yang for explanations of many of the biological concepts in this project.