LUND UNIVERSITY

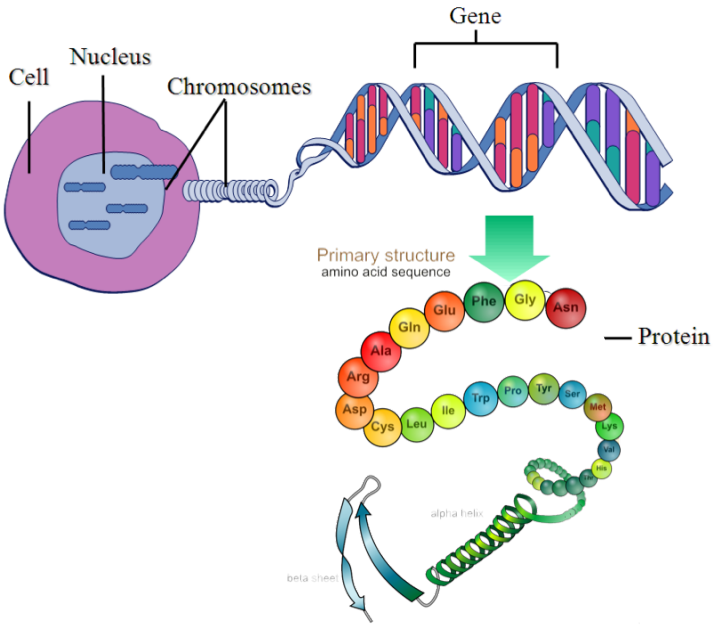
Personalized Medicine

Redefining Cancer Treatment

RAMONA BENDIAS, FRIDA BÖRNFORS

Is there a way to automatically classify genetic variations based on medical papers?

Biology crash course, pt 1



Images: http://www.aboutthemcat.org/images/organic-chemistry/primary-structure.png
https://www.diabetesqld.org.au/media-centre/2018/january/ study-reveals-new-diabetes-gene-in-families-with-rare-blood-glucose-conditions.aspx

LUND UNIVERSITY

Biology crash course, pt 2

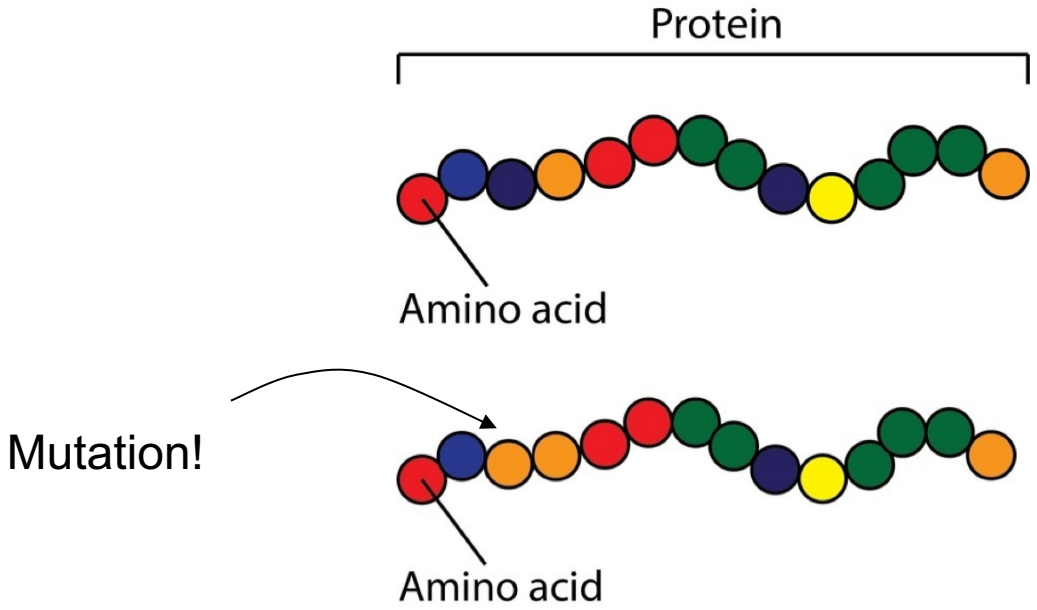


Image: https://www.quora.com/What-is-the-difference-between-amino-acids-and-enzymes

LUND UNIVERSITY

Biology crash course, pt 3

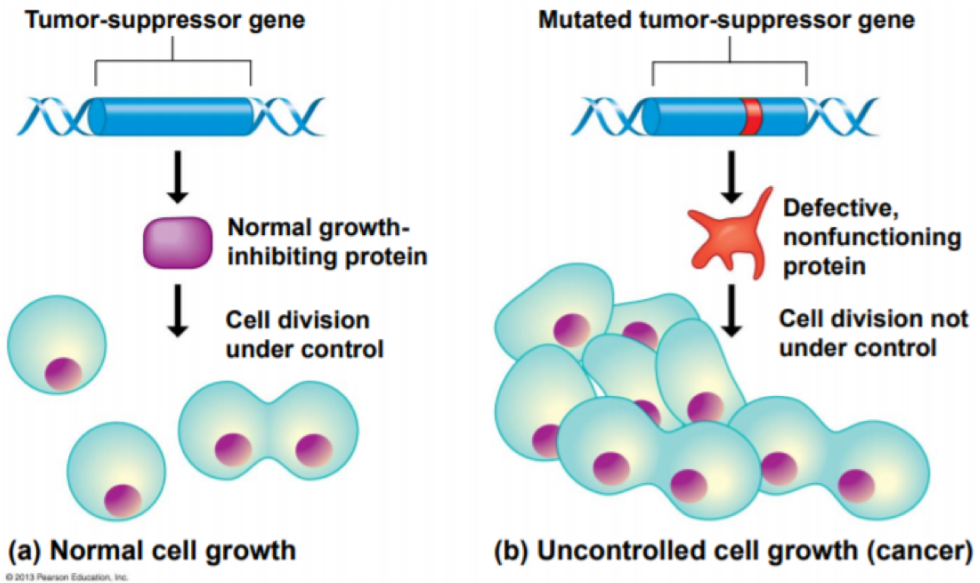


Image: Pearson Educational Inc

LUND UNIVERSITY

Data provided

Gene	Variation	Class	Text
FAM58A	Truncating Mutations	Likely Loss-of-function	Cyclin-dependent kinases (CDKs) regulate a var...
CBL	W802*	Likely Gain-of-function	Abstract Background Non-small cell lung canc...
CBL	Q249E	Likely Gain-of-function	Abstract Background Non-small cell lung canc...
CBL	N454D	Neutral	Recent evidence has demonstrated that acquired...
CBL	L399V	Loss-of-function	Oncogenic mutations in the monomeric Casitas B...
CBL	V391I	Loss-of-function	Oncogenic mutations in the monomeric Casitas B...
CBL	V430M	Likely Neutral	Oncogenic mutations in the monomeric Casitas B...
CBL	Deletion	Likely Loss-of-function	CBL is a negative regulator of activated recep...
CBL	Y371H	Loss-of-function	Abstract Juvenile myelomonocytic leukemia (JM...
CBL	C384R	Loss-of-function	Abstract Juvenile myelomonocytic leukemia (JM...

LUND UNIVERSITY

Example of how to classify

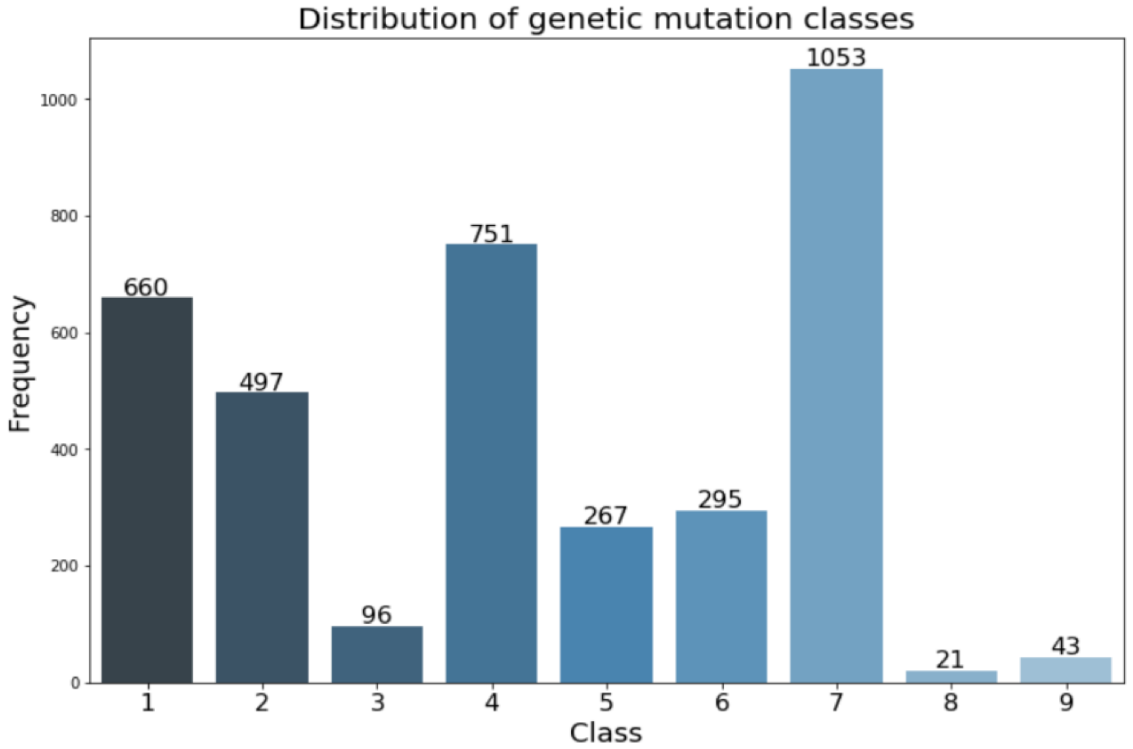
Gene	Variant	
CBL	V391I	[...] mutations (L399V, G375P, P395A and **V391I**) which **attenuated the CBL E3 activity**

→ Class 4, loss-of-function

Gene	Variant	
CBL	V430M	The second group of mutants (M374V, **V430M**, P428L, Q249E and double mutant S80N/H94Y) **maintained the CBL activity**

→ Class 5, likely neutral

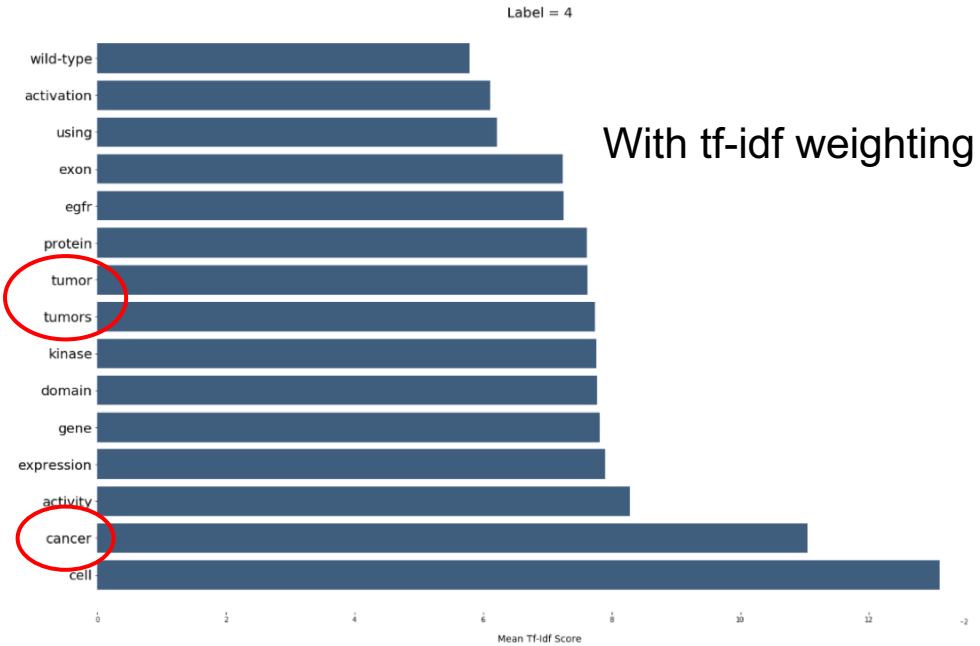
LUND UNIVERSITY



Total number of examples: 3683

LUND UNIVERSITY

Extract features from text



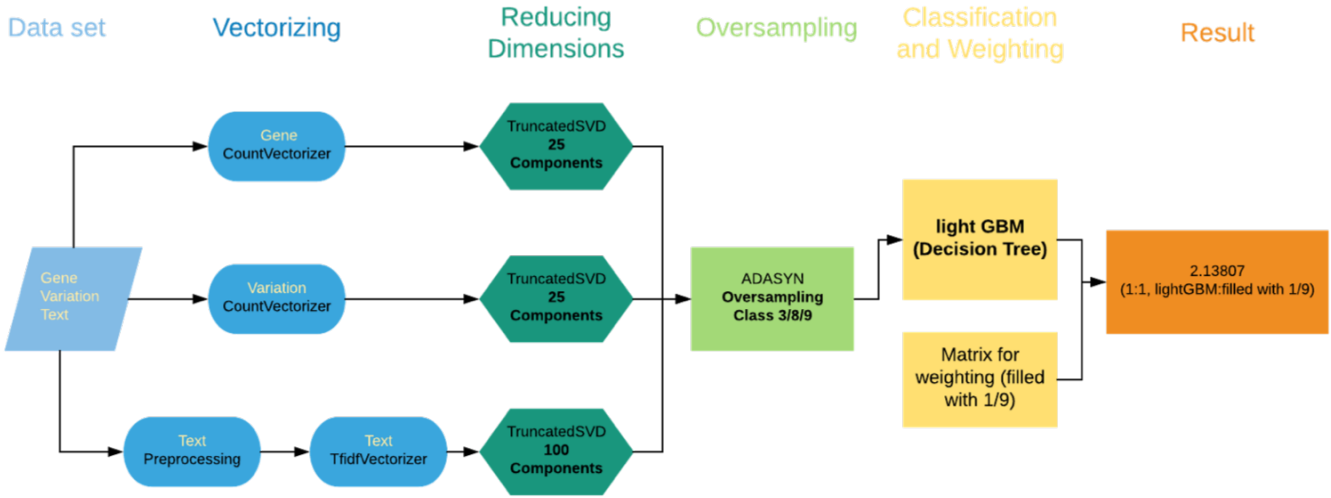
With tf-idf weighting

LUND UNIVERSITY

Include additional features

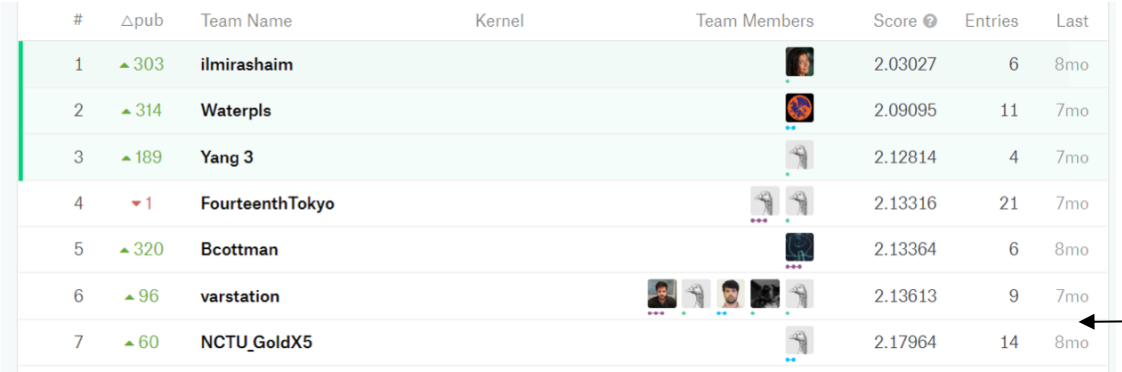
Variation: Q249E

Character n-grams
Q
Q2
Q24
Q249
Q249E
2
24
...



LUND UNIVERSITY

Results - Kaggle competition



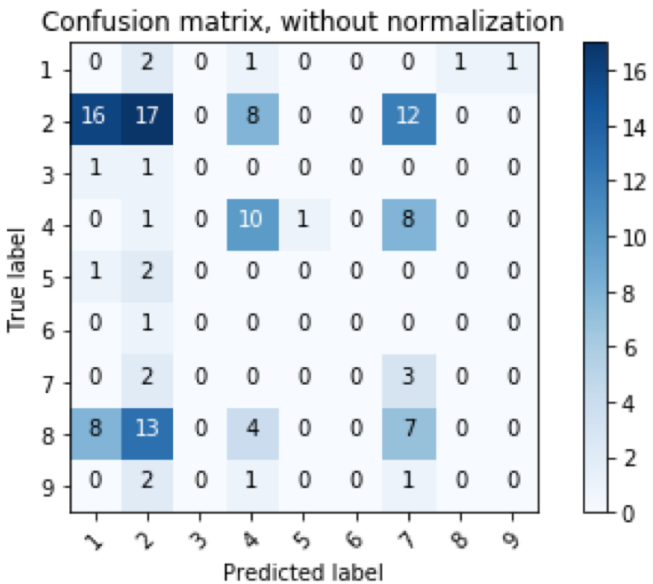
#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲303	ilmirashaim			2.03027	6	8mo
2	▲314	Waterpls			2.09095	11	7mo
3	▲189	Yang 3			2.12814	4	7mo
4	▼1	FourteenthTokyo			2.13316	21	7mo
5	▲320	Bcottman			2.13364	6	8mo
6	▲96	varstation			2.13613	9	7mo
7	▲60	NCTU_GoldX5			2.17964	14	8mo

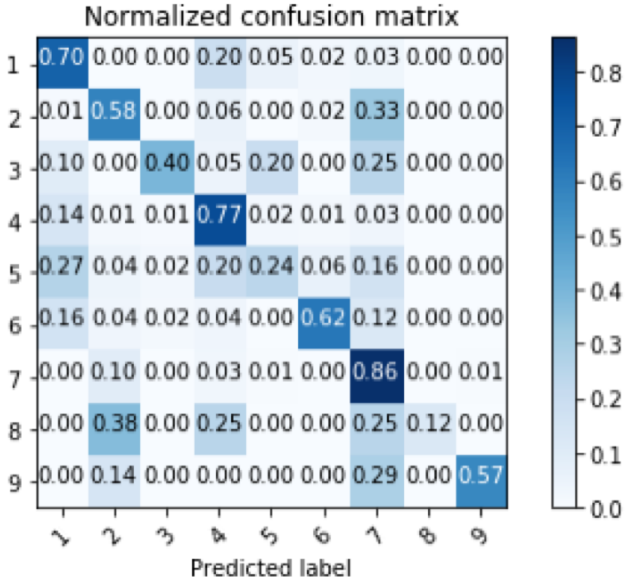
2.13807

Score = $$-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k}\log p_{i,k}$$

LUND UNIVERSITY

Confusion matrix





LUND UNIVERSITY

Reflections about the project

- Memory usage / Python
- Different approach
- Amount of data matters
- Challenging task
- Kaggle - good platform to learn machine learning!

Is there a way to automatically classify genetic variations based on medical papers?

Thank you for listening!

Questions?

LUND
UNIVERSITY