

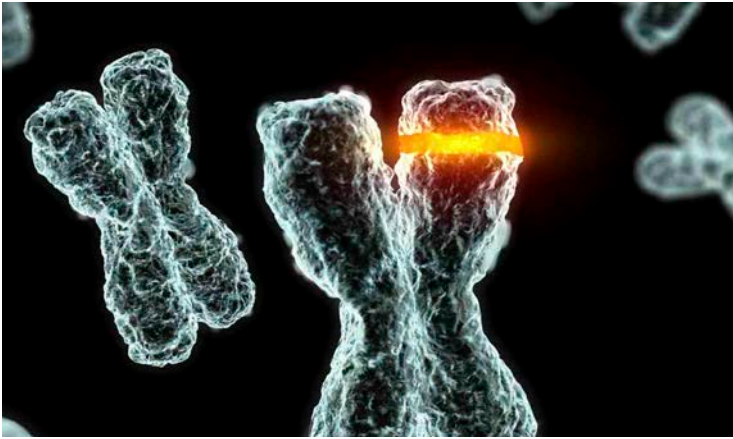


Personalized Medicine: Redefining Cancer Treatment

Team Karkinos



Introduction



Conventional Workflow

Step 1:

Select genetic mutations

Step 2:

Search in literature on selected mutations

Step 3:

Manually analyze text, and
classify mutations

tremendous time/efforts/expertise

Our Task

Task:

Automate Step 3

Dataset:

expert-annotated knowledge
base on genetic mutations

Classification problem:

Classify entries of **genes +
mutations + text** into 9 labeled
classes

Pipeline

Data & Preprocessing

Academic Articles

Shared Genetic Factors Involved in Celiac Disease, Type 2 Diabetes and Anorexia Nervosa Suggest Common Molecular Pathways for Chronic Diseases.
Stallanck, J. *et al.* *Cell* 2018; 174: 1105-1115. doi:10.1016/j.cell.2018.04.018

Abstract
BACKGROUND AND OBJECTIVES: Genome-wide association studies (GWAS) have identified several genetic variants associated with celiac disease, type 2 diabetes, and anorexia nervosa. These variants are enriched in over-representations of genes involved in type 2 diabetes and anorexia nervosa associated with celiac disease, suggesting involvement of common metabolic pathways for development of these chronic diseases. The aim of this study was to extend these previous analyses to study the gene expression in the gut from children with celiac disease.

RESULTS
Only six target genes involved in type 2 diabetes and four genes associated with anorexia nervosa were found to be differentially expressed in small intestinal biopsies from 144 children with celiac disease at median (range) age of 7.4 years (1.6-17.6) and from 154 disease controls at a median (range) age of 11.4 years (1.4-18.5).

CONCLUSIONS
Genetic factors involved in celiac disease, type 2 diabetes and anorexia nervosa suggest common underlying molecular pathways for these diseases.

Gene + Mutation Names



Feature Extraction

TF-IDF Text

CountVectorizer Gene
& Mutation n-grams

Latent Semantic
Analysis
Truncated SVD

Genes/mutation
families encoded

Word counts extracted
from text/genetic data

Modeling

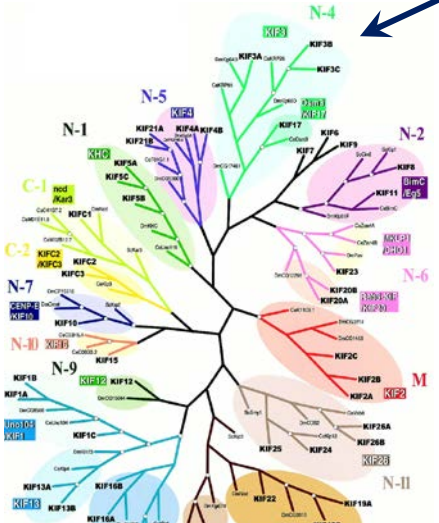
XGBoost modeling
+ cross validation

Ensemble of Top 5
XGBoost Models

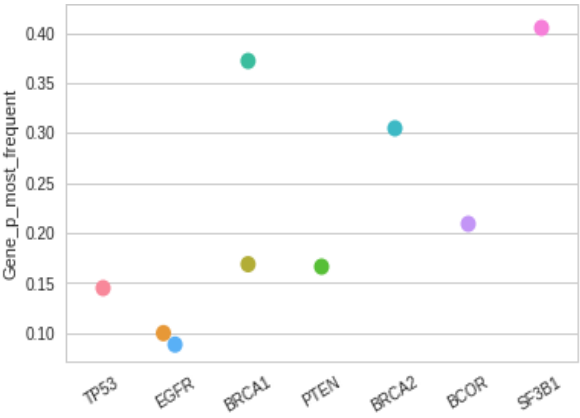
Input Data



Features Extracted & Unionized

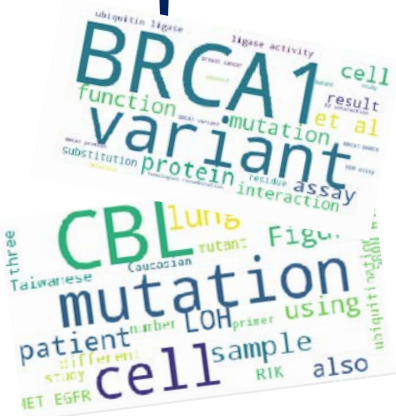


Gene Family Root & Mutation Types



Dimensions reduced with Truncated SVD

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9



Frequency of gene/mutation in text & word count

Tf-idf

| | Text1 | Text2 | Text... |
|--------|-------|-------|---------|
| cbl | | 0.87 | |
| cdk10 | 0.77 | | |
| cyclin | 0.25 | | |
| egfr | | 0.14 | |
| ets2 | 0.39 | | |
| fam58a | 0.26 | | |
| loh | | 0.12 | |
| lung | | 0.13 | |
| mutat | | 0.16 | |
| star | 0.12 | | |
| ... | | | |
| Wn | | | |

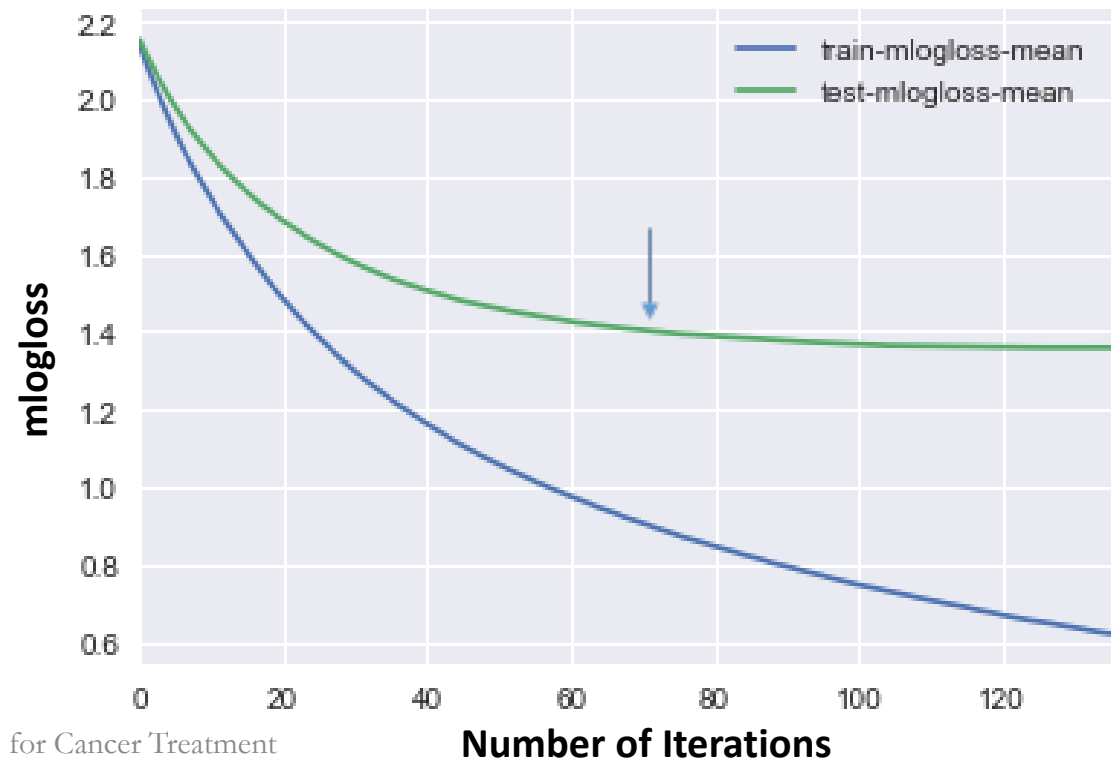
LSA (Truncated SVD)

Modeling

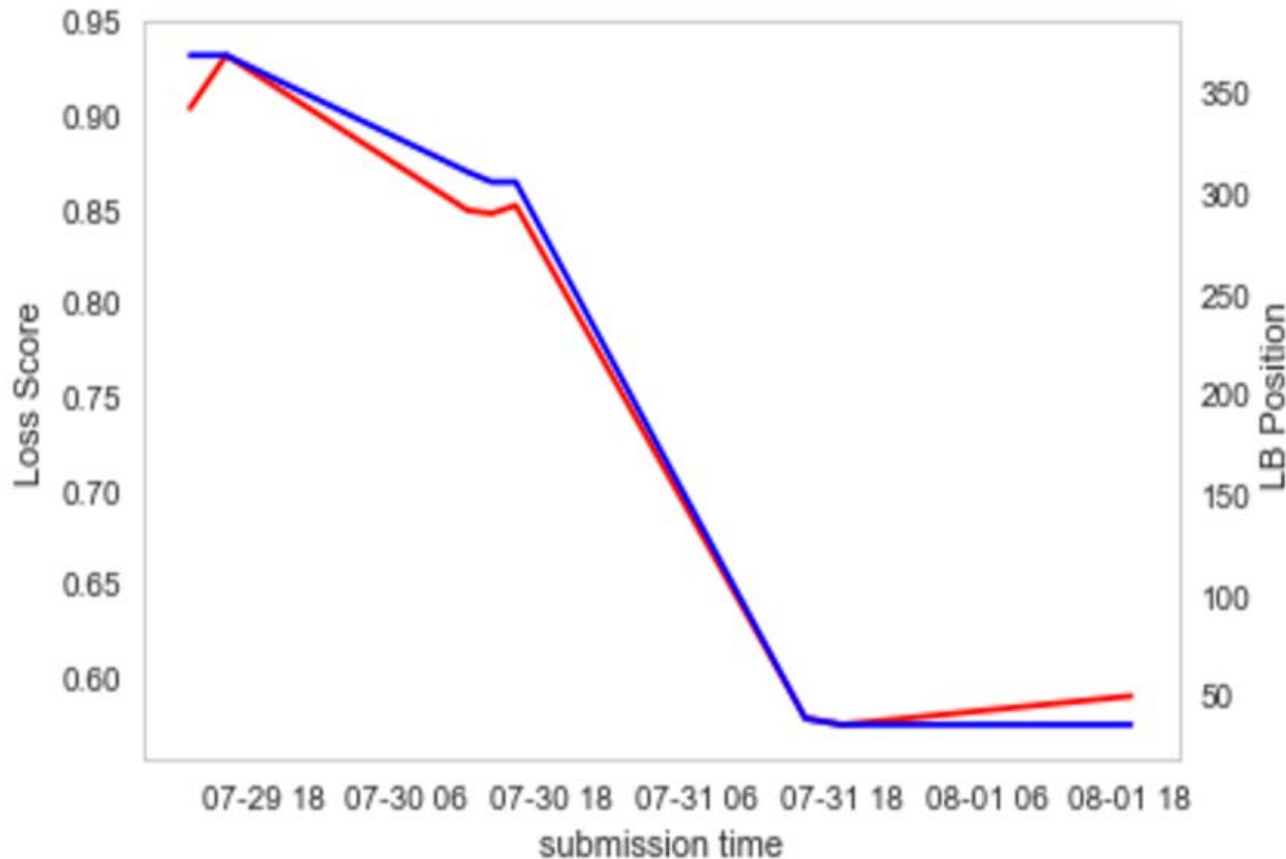
Model selection: **XGBoost**



| Cross Validation |
|---|
| Evaluation metric: mlogloss |
| Hyperparameter tuning: ensemble of 5 XGBoosts with top 5 sets of hyperparameter |
| Number of iterations: problem of overfitting |



Results & Future Steps



Final Log Score on Training: 0.29693

Final Log Score on Validation: 0.94718

Final Log Score on Testing: 0.58404

Final Leader-Board Position: Top 10%

Future Steps

1. Word2Vec model pre-trained on larger corpus of Bio texts for better text mining

2. Better ensembling of unrelated models and Stack generalization

Thank you!

