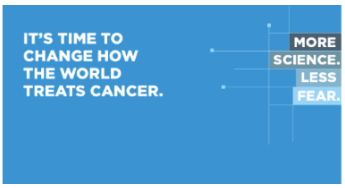

Personalized Medicine: Redefining Cancer Treatment

Predict the effect of Genetic Variants to enable Personalized Medicine

$15,000 · 449 teams · 2 months to go

Additional Files

- submissionFile
- test_text.zip
- test_variants.zip
- training_text.zip
- training_variants.zi...

training_variants.zip 24.25 KB Download

Data Introduction

In this competition you will develop algorithms to classify genetic mutations based on clinical evidence (text).

There are nine different classes a genetic mutation can be classified on.

This is not a trivial task since interpreting clinical evidence is very challenging even for human specialists. Therefore, modeling the clinical evidence (text) will be critical for the success of your approach.

Both, training and test, data sets are provided via two different files. One (training/test_variants) provides the information about the genetic mutations, whereas the other (training/test_text) provides the clinical evidence (text) that our human experts used to classify the genetic mutations. Both are linked via the ID field.

Therefore the genetic mutation (row) with ID=15 in the file training_variants, was classified using the clinical evidence (text) from the row with ID=15 in the file training_text

Finally, to make it more exciting!! Some of the test data is machine-generated to prevent hand labeling. You will submit all the results of your classification algorithm, and we will ignore the machine-generated samples.

File descriptions

- **training_variants** - a comma separated file containing the description of the genetic mutations used for training. Fields are **ID** (the id of the row used to link the mutation to the clinical evidence), **Gene** (the gene where this genetic mutation is located), **Variation** (the aminoacid change for this mutations), **Class** (1-9 the class this genetic mutation has been classified on)

- **training_text** - a double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations. Fields are **ID** (the id of the row used to link the clinical evidence to the genetic mutation), **Text** (the clinical evidence used to classify the genetic mutation)
- **test_variants** - a comma separated file containing the description of the genetic mutations used for training. Fields are **ID** (the id of the row used to link the mutation to the clinical evidence), **Gene** (the gene where this genetic mutation is located), **Variation** (the aminoacid change for this mutations)
- **test_text** - a double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations. Fields are **ID** (the id of the row used to link the clinical evidence to the genetic mutation), **Text** (the clinical evidence used to classify the genetic mutation)
- **submissionSample** - a sample submission file in the correct format