VEST (Variant Effect Scoring Tool) Developer Documentation

Downloading the VEST software package:

Download the software package and the SNVBox database here:
http://wiki.chasmsoftware.org/index.php/CHASM_DL

VEST consists of a number of python scripts and wrappers to produce the datasets needed for training a classifier and applying it to mutation data derived from genome sequencing. The PARF (Parallelized Random Forest) software is used for Random Forest Construction. VEST is distributed with a compiled version of PARF with minor modifications. The original PARF code is available from http://code.google.com/p/parf/.

Figure 1 VEST subdirectories

The VEST classifier is stored in the BuiltClassifiers directory under VEST. This directory includes the list of features used for classifier training and the set of null variants used to get empirical p-values.

Scripts used by SNVGet for interfacing with the SNVBox MySQL database.

Documentation.

Example datasets and command lines for classifier training and mutation scoring.

Scripts used by VEST for interfacing with MySQL.

Selecting which features to use to represent mutations:

The SNVBox stores 85 pre-computed features that can be used in classifier training. To change the list of features used in classifier training, create a new feature list that includes only the features of interest, 1 per line. The –f argument can be used to specify the custom feature list when training a classifier. By default, VEST will use all 85 features. The default list of 85 features can be found in the directory storing the VEST classifier distributed with the software package.

Modifying the VEST training set:

The disease class of the VEST training set is constructed from the Human Gene Mutation Database (HGMD) which requires a license and thus cannot be distributed with the VEST classifier. The neural mutation set is derived from variants detected in the Exome Sequencing Project http://evs.gs.washington.edu/EVS/ and consists of missense variants present in those samples at frequencies of 1% or higher. A new traininset can be constructed from a custom list of missense mutations using the SNVGet software. In general, the number of disease and neutral mutations used for classifier training should be balanced.

Adjusting Random Forest classifier training parameters:

Random Forest construction parameters (ntrees and mtry) can be adjusted in the vest_classifier.conf configuration file. A high mtry parameter will increase the chance of classifier over-fitting, by increasing the correlation between decision trees constituting the Random Forest.