CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) Developer Documentation

Downloading the CHASM software package:

Download the software package and the SNVBox database here:

http://wiki.chasmsoftware.org

Overview of CHASM software package

CHASM 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 tumor genome sequencing. The PARF (Parallelized Random Forest) software is used for Random Forest Construction. CHASM is distributed with a compiled version of PARF with minor modifications. The original PARF code is available from http://code.google.com/p/parf/.

Directory structure

CHASM	BuiltClassifiers	Pre-trained classifiers distributed with CHASM.
	ClassifierPack	Files used by CHASM for classifier construction and statistical analysis including the feature list and passenger rate tables.
	db	Scripts used by SNVGet for interfacing with the SNVBox MySQL database.
	doc	Documentation.
	example	Example datasets and command lines for classifier training and mutation scoring.
	util	Scripts used by CHASM for interfacing with MySQL as well as generating passenger mutations.

Using CHASM to develop your own driver mutation classifiers

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, CHASM will use all 85 features. Feature files are stored in the ClassifierPack directory in the features folder.

Modifying the CHASM training set
:

The driver class of the CHASM training set can be extended to include additional mutations, or filtered to remove mutations. The driver mutations are stored in CHASM/ClassifierPack/drivers.tmps. In general, the number of passenger mutations used for classifier training should be approximately matched to the number of drivers. The number of passengers to simulate is an adjustable parameter found in the chasm_classifiers.conf file.

Adjusting Random Forest classifier training parameters:

Random Forest construction parameters (ntrees and mtry) can be adjusted in the CHASM\classifiers.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.

Constructing custom passenger mutation rate table:

A passenger mutation rate table is created from tumor sequencing data by dividing up mutations based on neighboring DNA bases and nucleotide change. The passenger mutation rate table is a tab-delimited file with 5 rows and 9 columns. The first row gives column names. The first column indicates the alternative base that was substituted, and the subsequent columns represent the 8 di-nucleotide sequence categories (C in CpG (C*pG), G in sCpG (CpG*), C in TpC (TpC*), G in GpA (G*pA), A, C, T, G) into which base substitutions are divided for. The asterisk indicates the mutated base in the di-nucleotide contexts. Fields in the table that do not describe a nucleotide change (e.g. A to A, or CpG* to G) should be set to 0. All other fields contain the count of nucleotide substitutions where the mutated base is substituted with the base specified in the first column, normalized by the total number of base substitutions observed. Table 1 shows a correctly formatted passenger mutation rate table.

	C*pG	CpG*	TpC*	G*pA	A	C	G	T
-> A	0.009	0.227	0.014	0.031	0	0.043	0.077	0.028
-> C	0	0.005	0	0.015	0.012	0	0.021	0.028
-> G	0.005	0	0.009	0	0.060	0.029	0	0.015
-> T	0.172	0.003	0.021	0.023	0.025	0.074	0.057	0

Table 1 Passenger mutation rate table.

Passenger mutation rate table construction requires both bioinformatics expertise and cancer expertise. A set of pre-constructed passenger mutation rate tables are provided in (yourlocation)/CHASM/ClassifierPack/contexts/. If a new passenger mutation rate table is needed, the table can be constructed from tumor sequencing data as follows:

1. First gather the subset of base substitutions that result in nonsynonymous mutations when mapped onto proteins.

2. Next, remove any base substitutions that are known driver mutations or occur in frequently mutated genes (TP53, KRAS, PIK3CA, PTEN, CDKN2A, SMAD4, NF1, RB1. etc). Some genes may be tumor specific, for example, cMET in head and neck cancer.

3. Group all remaining mutations into 8 categories (C in CpG (C*pG), G in CpG (CpG*), C in TpC (TpC*) , G in GpA (G*pA), A, C, T, G) based on neighboring DNA bases (see Figure 3). The categories are listed in order such that a mutation of TCG to TAG would fall into the C*pG category (since the mutated C is followed by a G) rather than the TpC* category or the C category.

4. Further divide each of the mutations falling into each of the 8 categories according to the base that was substituted and count the number of mutations.

5. Normalize the mutation counts by the total number of nonsynonymous mutations.

6. Format the normalized counts as shown in Table 1.

Pre-constructed passenger mutation rate tables for a variety of tumor types are included in the ClassifierPack directory in the context folder.