SNVBox
Basic User Documentation

Downloading the SNVBox database

The SNVBox database and SNVGet python script are available here:

http://wiki.chasmsoftware.org/index.php/Download

Installing SNVBox:

Requirements:

- Linux operating system (32bit/64bit).
- At least 60 GB disk space.
- 1-2 GB RAM.
- MySQL server 5.0 or newer.
- Python MySQLdb module.

To install the SNVBox database and SNVGet software package:

- Change directory to the location where you will install SNVGet with cd (yourlocation)
- Download the SNVGet install pack for the latest release and move it to the selected directory. Extract the files using tar xvzf SNVBOXSRC.tar.gz
- Download the latest release of the SNVBox MySQL database. This is a large file.
- Use the following commands from within MySQL to prepare for loading the database:

    - create database SNVBOX
    - grant ALL Privileges on SNVBOX.* to chasm_user@localhost identified by `dfltPass!';
    - grant RELOAD on *.* to chasm_user@localhost;

- Load the database into MySQL. At the command line run gunzip < SNVBox.sql.gz | mysql -u chasm_user -p SNVBOX and enter the password dfltPass! when prompted.
- Modify the snv_box.conf configuration files if necessary.

To use SNVGet:

SNVGet will retrieve feature values for a list of mutations. SNVGet therefore expects a file with a list of feature names (1 per line) as well as the list of mutations for which to retrieve features. SNVGet also requires the user to specify a name for the output file.

SNVGet commandline arguments:

-f specify file with list of features to retrieve
-c include class labels – specify separate mutation lists for each class
-o output file name
-r return raw SNVBox feature values. By default values are scaled by mean and rms.
-m allow missing values (specified by 'NA'). By default values are filled with mean.

Formatting your list of mutations:

Mutations can be specified as tab delimited lists of transcript and amino acid substitution with codon number ('transcript coordinates') or as genomic location and nucleotide substitution ('genomic coordinates'). The expected format for each coordinate system is described here:

Transcript coordinates:

- (optional) mutation identifier
- transcript
- amino acid substitution (e.g. G12D or H1047R)

Example of transcript coordinates:

```
# UID / Transcript / AA change
TR1   NM_001126116.1   D127Y
TR2   NM_001144919.1   R162Q
TR3   NM_000321.2      Q702K
TR4   NM_000245.2      A1108S
TR5   NM_004333.4      V600E
TR6   NM_001005862.1   G746V
```

Genomic Coordinates:

- (optional) mutation identifier
- chromosome (e.g. "chr1")
- 0-based start position of the nucleotide
- 1-based end position of the nucleotide
- strand of the nucleotides being reported in the following column*
- reference nucleotide
- alternative nucleotide

*If the reference and alternative nucleotides match the forward strand of the reference genome this should be "+", if they match the reverse strand it should be "-"

Example of genomic coordinates:

```
# UID / Chr. / Start / Stop / Strand / Ref. base / Alt. base
TR1   chr17  7577505    7577506    -   G   T
TR2   chr10  123279679  123279680  -   G   A
TR3   chr13  49033966   49033967   +   C   A
TR4   chr7   116417504  116417505  +   G   T
TR5   chr7   140453135  140453136  -   T   A
TR6   chr17  37880997   37880998   +   G   T
```

Retrieving features with SNVGet:

If mutations are in transcript coordinates, SNVGetTranscript should be run to retrieve features:

./snvGetTranscript –f featurelist –o outputfilename.arff mutationlist

or if mutations are in genomic coordinates:

./snvGetGenomic –f featurelist –o outputfilename.arff mutationlist

The –c option allows the user to specify class labels to be included in the output file. The user needs to specify a separate file of mutations for each class.

./snvGetTranscript –c –f featurelist –o outputfilename.arff class1label mutationlist1 class2label mutationlist2

SNVGet will generate an arff file with the list of features. Please note that the arff file generated is not quite a correctly formatted arff. This is because SNVGet was intended to run with the PARF (PArallel Random Forsest) software that requires a special variant of the arff format. See http://code.google.com/p/parf/wiki/DatasetFileFormat for details.

Features available in SNVBox

SnvGet Feature Name	Description
AABLOSUM	Amino acid substitution score from the BLOSUM 62 matrix
AACharge	Change in formal charge resulting from replacing the reference amino acid residue with the mutation. Histidine is assumed protonated (formal charge of +1).
AACOSMIC	Ln(frequency) of missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) in COSMIC (release 38)
AACOSMICvsHapMap	Ln(frequency) of missense change in COSMIC (release 38) normalized by the number of times the change type was observed in HapMap validated SNPs in dbSNP Build 129
AACOSMICvsSWISSPROT	Ln(frequency) of missense change in COSMIC (release 38) normalized by the frequency of reference amino acid residue in human proteins in SwissProt/TrEMBL
AAEx	Amino acid substitution score from the EX matrix
AAGrantham	The Grantham distance from reference to mutation amino acid residue
AAHapMap	Ln(frequency) of missense change type in HapMap validated SNPs in dbSNP Build 129
AAHGMD2003	Number of times that the reference to mutation substitution occurs in the Human Gene Mutation Database, 2003 version
AAHydrophobicity	The change in hydrophobicity resulting from the substitution
AAMJ	Amino acid substitution score from the Miyazawa-Jernigan contact energy matrix
AAPAM250	Amino acid substitution score from the PAM250 matrix
AAPolarity	Polarity change from reference to mutation amino acid residue
AATransition	Frequency of transition between two neighboring amino acids based on all human proteins in SwissProt/TrEMBL
AATripletFirstDiffProb	Difference in probability of occurrence of reference and mutation amino acid residue in the 1st position
AATripletFirstProbMut	Probability of seeing the mutant amino acid in position 1 of a triplet
AATripletFirstProbWild	Probability of seeing the wild-type amino acid in position 1 of a triplet
AATripletSecondDiffProb	Difference in probability of occurrence of reference and mutation amino acid residue in the 2nd position
AATripletSecondProbMut	Probability of seeing the mutant amino acid in position 2 of a triplet
AATripletSecondProbWild	Probability of seeing the wild-type amino acid in position 2 of a triplet
AATripletThirdDiffProb	Difference in probability of occurrence of reference and mutation amino acid residue in the 3rd position
AATripletThirdProbMut	Probability of seeing the mutant amino acid in position 3 of a triplet
AATripletThirdProbWild	Probability of seeing the wild-type amino acid in position 3 of a triplet
AAVB	Amino acid substitution score from the VB (Venkatarajan and Braun) matrix
AAVolume	Change in residue volume resulting from the replacement (in units of cubic Angstroms)
ExonConservation	Conservation score for the entire exon calculated from a 46-species phylogenetic alignment using the UCSC Genome Browser (hg19). Scores are given for windows of nucleotides. Retrieved are the scores for each region that overlaps the exon in which the base substitution occurred, and a weighted average of the conservation scores is calculated, where the weight is the number of bases with a particular score.
ExonHapMapSnpDensity	Number of HapMap verified SNPs (dbSNP build 131) in the

	exon where the mutation is located divided by the length of the exon.
ExonSnpDensity	Number of SNPs in the exon where the mutation is located divided by the length of the exon.
HMMEntropy	Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment, corresponding to the location of the mutation.
HMMPHC	Measure which is calculated based on the degree of conservation of the residue, the mutation and the most probable amino acid in a match state of a hidden Markov model built with SAM-T2K software
HMMRelEntropy	Kullback-Leibler Divergence calculated for the column of the SAM-T2K multiple sequence alignment (corresponding to the location of the mutation) and that of a background distribution of amino acid residues computed from a large sample of multiple sequence alignments.
MGAEntropy	Shannon entropy calculated for the column of the Multiz-46-way alignment, corresponding to the location of the mutation.
MGAPHC	Measure which is calculated based on the degree of conservation of the residue, the mutation and the most probable amino acid in the column of a Multiz-46-way alignment from UCSC Human Genome Browser hg19
MGARelEntropy	Kullback-Leibler divergence calculated for the column of Multiz-46-way alignment (corresponding to the location of the mutation) and that of a background distribution of amino acid residues computed from a large sample of multiple sequence alignments
PredBFactorF	Probability that the wild type residue backbone is flexible
PredBFactorM	Probability that the wild type residue backbone is intermediatly flexible
PredBFactorS	Probability that the wild type residue backbone is stiff
PredRSAB	Probability of the wild type accessibility residue being buried
PredRSAE	Probability of the wild type accessibility residue being exposed
PredRSAI	Probability of the wild type accessibility residue being intermediatly exposed
PredSSC	Probability that the secondary structure of the region in which the wild type residue exists is loop
PredSSE	Probability that the secondary structure of the region in which the wild type residue exists is strand
PredSSH	Probability that the secondary structure of the region in which the wild type residue exists is helix
PredStabilityH	Probability that the wild stability type residue contributes to overall protein stability in a manner that is highly stabilizing. Stability estimates for the neural network training data were calculated using the FoldX force field.
PredStabilityL	Probability that the wild stability type residue contributes to overall protein stability in a manner that is average. Stability estimates for the neural network training data were calculated using the FoldX force field
PredStabilityM	Probability that the wild stability type residue contributes to overall protein stability in a manner that is destabilizing, Stability estimates for the neural network training data were calculated using the FoldX force field
RegCompC	Proportion of Cysteines around position
RegCompDE	Proportion of Aspartic and Glutamic Acids around position
RegCompEntropy	Shannon entropy of amino acid residues around position
RegCompG	Proportion of Glycines around position
RegCompH	Proportion of Histidines around position
RegCompILVM	Proportion of Isoleucines, Leucines, Valines, and Methionines around position
RegCompKR	Proportion of Lysines and Arginines around position

RegCompNormEntropy	Shannon entropy of amino acid residues around position normalized by the number of different amino acids within the window
RegCompP	Proportion of Prolines around position
RegCompQ	Proportion of Glutamines around position
RegCompWYF	Proportion of Tryptophans, Tyrosines, and Phenylalanines around position
UniprotACTSITE	Sites involved in enzymatic activity
UniprotBINDING	Binding sites. 1 indicates its presence and 0 absence.
UniprotCABIND	Calcium binding site. 1 indicates its presence and 0 absence.
UniprotCARBOHYD	Carbohydrate binding site. 1 indicates its presence and 0 absence.
UniprotCOMPBIAS	Compositionally biased region. 1 indicates its presence and 0 absence.
UniprotDISULFID	Site of disulfide bond. 1 indicates its presence and 0 absence.
UniprotDNABIND	DNA binding site. 1 indicates its presence and 0 absence.
UniprotDOM_Chrom	Site in a domain involved in chromatin structure remodeling. 1 indicates its presence and 0 absence.
UniprotDOM_LOC	Site in a domain that determines correct cellular localization of a protein. 1 indicates its presence and 0 absence.-
UniprotDOM_MMBRBD	Site in a domain that binds to the cell membrane. 1 indicates its presence and 0 absence.
UniprotDOM_PostModEnz	Site in an enzymatic domain responsible for any kind of post-translational modification. 1 indicates its presence and 0 absence.
UniprotDOM_PostModRec	Site in a domain that recognizes a post-translationally modified residue. 1 indicates its presence and 0 absence.
UniprotDOM_PPI	Site in a protein-protein interaction domain. 1 indicates its presence and 0 absence.
UniprotDOM_RNABD	Site in an RNA binding domain. 1 indicates its presence and 0 absence.
UniprotDOM_TF	Site in a transcription factor domain. 1 indicates its presence and 0 absence.
UniprotLIPID	Lipid binding site. 1 indicates its presence and 0 absence.
UniprotMETAL	Metal binding site. 1 indicates its presence and 0 absence.
UniprotMODRES	Site of modified residue. 1 indicates its presence and 0 absence.
UniprotMOTIF	Site of known functional motif. 1 indicates its presence and 0 absence.
UniprotNPBIND	Nucleotide phosphate-binding region. 1 indicates its presence and 0 absence.
UniprotPROPEP	Site in the propeptide (cleaved in mature protein). 1 indicates its presence and 0 absence.
UniprotREGIONS	Region of interest in the protein sequence. 1 indicates its presence and 0 absence.
UniprotREP	Repeat region. 1 indicates its presence and 0 absence.
UniprotSECYS	Site of selenocysteine. 1 indicates its presence and 0 absence.
UniprotSIGNAL	Site of localization signal (protein targeted to secretory pathway or periplasm). 1 indicates its presence and 0 absence.
UniprotSITE	An interesting amino acid site in the protein sequence. 1 indicates its presence and 0 absence.
UniprotTRANSMEM	Transmembrane region. 1 indicates its presence and 0 absence.
UniprotZNFINGER	Site in a zinc finger. 1 indicates its presence and 0 absence.

For additional details, please use the CHASM wiki http://wiki.chamsoftware.org