CHASM, VEST, and SNVBox Feature Retrieval Tool User Manual

Last updated: 4th March 2014

1. CHASM

1.1 Training the Classifier

1.1.1 Cancer-specific mutation context table.

This method requires a custom passenger mutation rate table that specifies the frequency of nucleotide changes in different DNA sequence contexts for the cancer you are studying. We provide passenger mutation rate tables for several tissues in the ClassifierPack. A passenger mutation rate table itself is a tab delimited text file in the following format:

	C*pG	CpG*	TpC*	G*pA	A	C	G	T
->A	1.01	5.36	0.38	0.55	0.00	0.59	0.50	0.21
->C	0.00	0.64	0.00	0.52	0.02	0.00	0.32	0.36
->G	0.64	0.00	0.43	0.00	0.30	0.32	0.00	0.18
->T	5.79	0.98	0.61	0.32	0.21	0.59	0.58	0.00

The rates are the counts of mutations in each DNA sequence category normalized by the total number of nonsynonymous mutations observed. The asterisk indicates the mutated base in the di-nucleotide contexts. Please note that the mutation contexts are non-overlapping, which means that if a C to G mutation is observed in a C*pG context, the same mutation won't be counted for both the TpC* and C contexts.

1.1.2. Building your Classifier

Run BuildClassifier in the (Installation directory)/CHASM:

```
>BuildClassifier –m MutationTable -o ClassifierName -f FeaturesList
```

where MutationTable is the location of the Category-specific mutation context table generated in step (1), and ClassifierName is the name of the classifier. All generated classifiers are contained in the (installation directory)/BuiltClassifiers folder.

If mutations to be scored by CHASM are in genomic coordinates, the BuildClassifier script should be run with the "-g" option.

Note that "-f FeaturesList" is optional. If it is not specified, the default feature list will be used.

1.2 Formatting Mutation Data

Lists of mutations to score should be provided in tab-delimited files with no header. The classifier accepts mutations in 2 coordinates:

1.2.1 Transcript coordinates

- Without mutation UID: Transcript <tab or space> Amino acid substitution

```
NP_001135977 R641W
NP_835455 R151C
NP_055645 L590V
NP_689808 D28H
NP_005472 S372R
NP_112493 S35R
NP_859061 A118V
NP_892018 R153C
NP_001074003 R264Q
NP_001073893 R1272C
```

- With mutation UID: Mutation UID <tab or space> Transcript <tab or space> Amino acid substitution

```
1 NP_001135977 R641W
2 NP_835455 R151C
3 NP_055645 L590V
4 NP_689808 D28H
5 NP_005472 S372R
6 NP_112493 S35R
7 NP_859061 A118V
8 NP_892018 R153C
9 NP_001074003 R264Q
10 NP_001073893 R1272C
```

Currently SNVBox supports feature retrieval for Refseq, CCDS, and Ensembl accessions.

1.2.2. Genomic coordinates

- Without mutation UID: Chromosome <tab or space> chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
chr22  25115449     +      A      G
chr22  25119120     +      A      C
chr22  25124311     ---    C      G
chr22  25144912     +      C      T
chr22  25145753     ---    C      T
chr22  25147423     +      T      A
chr22  25150138     +      A      G
chr22  25152618     +      C      T
chr22  25158438     +      C      T
chr22  24121378     +      G      T
```

- With User assigned ID: 7 column format: chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
1      chr22  25115449     +      A      G
```

```
2     chr22  25119120    +     A     C
3     chr22  25124311    ---   C     G
4     chr22  25144912    +     C     T
5     chr22  25145753    ---   C     T
6     chr22  25147423    +     T     A
7     chr22  25150138    +     A     G
8     chr22  25152618    +     C     T
9     chr22  25158438    +     C     T
10    chr22  24121378    +     G     T
```

Important: Coordinates are expected to be on the GRCh37/hg19 build of the human genome.

If no mutation ID is assigned, CHASM will automatically assign the row number of the mutation in the input file as the ID. The ID is useful for matching CHASM scores to mutations in the original input file.

1.3 Running the Classifier

1.3.1. Running the classifier

Run RunChasm in the CHASM_classifiers source directory as follows:

```
>./RunChasm classifier_name mutation_list
```

where classifier_name is the name of the classifier you built earlier, and mutation_list is the location of the list of mutations generated in step (1).

If mutations are in genomic coordinates, the RunChasm script should be run with the –g option.

1.3.2. Output

The following files will be generated in the directory containing your mutation list file:

- <input file name>.arff: The ARFF (Attribute-Relation File Format) file generated by SNV-Box, which has the following format.

```
@relation headerfile
@attribute UID string
@attribute ID string
@attribute ExonConservation numeric
@attribute ExonSnpDensity numeric
@attribute ExonHapMapSnpDensity numeric
@attribute HMMRelEntropy numeric
@attribute HMMEntropy numeric
@attribute HMMPHC numeric
@attribute MGARelEntropy numeric
@attribute MGAEntropy numeric
@attribute MGAPHC numeric
...
@data
1 NP_955533_P937A 0.693567650685 0.00588235294118 0.0 0.191425 ...
2 NP_056193_A1412V 0.763530500574 0.00496838301716 0.000451671183379 ...
```

```
3 NP_075390_L122V 0.520834494575 0.0448717948718 0.00961538461538 …
4 NP_001009611_Y459C 0.540801628658 0.0838926174497 0.00335570469799 ...
```

- <input file name>.output: Tab delimited text file with the following information.

  1st Column: Mutation ID (row number is used if no ID was given in the input file)
  2nd Column: Transcript_amino acid change
  3rd Column: Raw CHASM score
  4th Column: P-value
  5th Column: Benjamini-Hochberg false discovery rate

```
1      NP_955533_P937A    0.916 0.871 0.91
2      NP_056193_A1412V   0.652 0.175 0.78
3      NP_075390_L122V    0.562 0.103 0.78
4      NP_001009611_Y459C 0.408 0.041 0.78
5      NP_001009611_L454F 0.352 0.029 0.78
6      NP_001009611_F380I 0.314 0.021 0.78
7      NP_001009611_R144K 0.414 0.043 0.78
```

1.4 Error Messages

Please note that you may encounter SNVGet error messages (see chapter 3.4) while running CHASM during feature retrieval.

2. VEST

2.1 Formatting Mutation Data

Lists of mutations to score should be provided in tab-delimited files with no header. The classifier accepts mutations in 2 coordinates.

2.1.1 Transcript coordinates

- Without mutation UID: Transcript <tab or space> Amino acid substitution

```
NP_001135977 R641W
NP_835455 R151C
NP_055645 L590V
NP_689808 D28H
NP_005472 S372R
NP_112493 S35R
NP_859061 A118V
NP_892018 R153C
NP_001074003 R264Q
NP_001073893 R1272C
```

- With mutation UID: Mutation UID <tab or space> Transcript <tab or space> Amino acid substitution

```
1 NP_001135977 R641W
2 NP_835455 R151C
3 NP_055645 L590V
4 NP_689808 D28H
```

```
5 NP_005472 S372R
6 NP_112493 S35R
7 NP_859061 A118V
8 NP_892018 R153C
9 NP_001074003 R264Q
10 NP_001073893 R1272C
```

Currently SNVBox supports feature retrieval for Refseq, CCDS, and Ensembl accessions.

2.1.2. Genomic coordinates

- Without mutation UID: Chromosome <tab or space> chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
chr22  25115449     +      A      G
chr22  25119120     +      A      C
chr22  25124311     ---    C      G
chr22  25144912     +      C      T
chr22  25145753     ---    C      T
chr22  25147423     +      T      A
chr22  25150138     +      A      G
chr22  25152618     +      C      T
chr22  25158438     +      C      T
chr22  24121378     +      G      T
```

- With User assigned ID: 7 column format: chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
1      chr22  25115449     +      A      G
2      chr22  25119120     +      A      C
3      chr22  25124311     ---    C      G
4      chr22  25144912     +      C      T
5      chr22  25145753     ---    C      T
6      chr22  25147423     +      T      A
7      chr22  25150138     +      A      G
8      chr22  25152618     +      C      T
9      chr22  25158438     +      C      T
10     chr22  24121378     +      G      T
```

Important: Coordinates are expected to be on the GRCh37/hg19 build of the human genome.

If no mutation ID is assigned, VEST will automatically assign the row number of the mutation in the input file as the ID. The ID is useful for matching VEST scores to mutations in the original input file.

2.2 Running the Classifier

2.2.1. Running the classifier

Please run RunVest in the VEST_classifier source directory:

```
>./RunVest mutation_list -c classifier_name
```

where the “-c” argument is optional and should only be used if classifier_name is other than the default (VEST), and mutation_list is the location of the list of mutations.

If mutations are in genomic coordinates, the RunVest script should be run with the “-g” option.

2.2.2 Output

- <input file name>.arff: ARFF (Attribute-Relation File Format) file generated by SNV-Box

  ```
  @relation headerfile
  @attribute UID string
  @attribute ID string
  @attribute ExonConservation numeric
  @attribute ExonSnpDensity numeric
  @attribute ExonHapMapSnpDensity numeric
  @attribute HMMRelEntropy numeric
  @attribute HMMEntropy numeric
  @attribute HMMPHC numeric
  @attribute MGARelEntropy numeric
  @attribute MGAEntropy numeric
  @attribute MGAPHC numeric
  ...
  @data
  1 NP_955533_P937A 0.693567650685 0.00588235294118 0.0 0.191425 ...
  2 NP_056193_A1412V 0.763530500574 0.00496838301716 0.000451671183379 ...
  3 NP_075390_L122V 0.520834494575 0.0448717948718 0.00961538461538 ...
  4 NP_001009611_Y459C 0.540801628658 0.0838926174497 0.00335570469799 ...
  ```

- <input file name>.output: Tab delimited text file with the following information.

  1st Column: Mutation ID (row number is used if no ID was given)
  2nd Column: Transcript_amino acid change
  3rd Column: Raw VEST score
  4th Column: P-value
  5th Column: Benjamini-Hochberg false discovery rate (shown only if more than 10 mutations are scored)

  ```
  1 NP_955533_P937A 0.2690 0.4109
  2 NP_056193_A1412V 0.1340 0.7140
  3 NP_075390_L122V  0.0530 0.9411
  4 NP_001009611_Y459C 0.3550 0.2884
  5 NP_001009611_L454F 0.3100 0.3479
  6 NP_001009611_F380I  0.1590 0.6508
  7 NP_001009611_R144K 0.1080 0.7958
  ```

2.3 Error Messages

Please note that you may encounter SNVGet error messages (see chapter 3.4) while running VEST during

feature retrieval.

3. SNVGet: SNV-Box Feature Retrieval Tool

There are two tools for retrieving SNVBox features, snvGetGenomic and snvGetTranscript, for genomic and transcript-coordinate mutations.

3.1 Formatting Mutation Data

Lists of mutations to score should be provided in tab-delimited files with no header. The classifier accepts mutations in 2 formats:

3.1.1 Transcript coordinates

- Without mutation UID: Transcript <tab or space> Amino acid substitution

```
NP_001135977 R641W
NP_835455 R151C
NP_055645 L590V
NP_689808 D28H
NP_005472 S372R
NP_112493 S35R
NP_859061 A118V
NP_892018 R153C
NP_001074003 R264Q
NP_001073893 R1272C
```

- With mutation UID: Mutation UID <tab or space> Transcript <tab or space> Amino acid substitution

```
1 NP_001135977 R641W
2 NP_835455 R151C
3 NP_055645 L590V
4 NP_689808 D28H
5 NP_005472 S372R
6 NP_112493 S35R
7 NP_859061 A118V
8 NP_892018 R153C
9 NP_001074003 R264Q
10 NP_001073893 R1272C
```

Currently SNVBox supports feature retrieval for Refseq, CCDS, and Ensembl accessions.

3.1.2. Genomic coordinates

- Without mutation UID: Chromosome <tab or space> chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
chr22  25115449     +      A      G
chr22  25119120     +      A      C
chr22  25124311     ---    C      G
chr22  25144912     +      C      T
```

```
chr22  25145753     ---    C      T
chr22  25147423     +      T      A
chr22  25150138     +      A      G
chr22  25152618     +      C      T
chr22  25158438     +      C      T
chr22  24121378     +      G      T
```

- With User assigned ID: 7 column format: chromosome <tab or space> 1-based position <tab or space> strand on which reference and mutation bases are reported <tab or space> reference base <tab or space> mutation base

```
1      chr22  25115449     +      A      G
2      chr22  25119120     +      A      C
3      chr22  25124311     ---    C      G
4      chr22  25144912     +      C      T
5      chr22  25145753     ---    C      T
6      chr22  25147423     +      T      A
7      chr22  25150138     +      A      G
8      chr22  25152618     +      C      T
9      chr22  25158438     +      C      T
10     chr22  24121378     +      G      T
```

Important: Coordinates are expected to be on the GRCh37/hg19 build of the human genome.

If no mutation ID is assigned, SNVGet will automatically assign the row number of the mutation in the input file as the ID. The ID is useful for matching SNVBox feature values to mutations in the original input file.

3.2. Custom Features

SNV-Box accepts a list of custom features. Prepare a text file with the list of features you want:

```
ExonConservation
ExonSnpDensity
ExonHapMapSn
Density
HMMRelEntropy
HMMEntropy
HMMPHC
MGARelEntropy
MGAEntropy
MGAPHC
PredRSAB
PredRSAE
PredBFactorF
PredBFactorS
PredSSC
PredSSE
PredSSH
AAHydrophobicity
AAVolume
AAPolarity
...
```

3.3 Retrieving Features

3.3.1 Transcript coordinates

- To retrieve features from a single mutation file:

  ```
  >./snvGetTranscript -f [feature list file] -o [output arff file]
  [mutation file]
  ```

- To retrieve features from multiple classes the file names as class labels:

  ```
  >./snvGetTranscript -f [feature list file] -o [output arff file]
  [mutation file 1 as class 1] [mutation file 2 as class 2] etc.
  ```

- To retrieve features from multiple classes with custom class labels:

  ```
  >.snvGetTranscript -c -f [feature list file] -o [output arff file]
  [class 1] [mutation file 1] [class 2] [mutation file 2] etc.
  ```

3.3.2 Genomic coordinates

- To retrieve features from a single mutation file:

  ```
  >./snvGetGenomic -f [feature list file] -o [output arff file] [mutation
  file]
  ```

- To retrieve features from multiple classes the file names as class labels:

  ```
  >./snvGetGenomic -f [feature list file] -o [output arff file]
  [mutation file 1 as class 1] [mutation file 2 as class 2] etc.
  ```

- To retrieve features from multiple classes with custom class labels:

  ```
  >./snvGetGenomic -c -f [feature list file] -o [output arff file] [class
  1] [mutation file 1] [class 2] [mutation file 2] etc.
  ```

- To retrieve transcript names for genomic coordinates:

  ```
  >./snvGetTranscriptList -f [mutation file] –c (optional – will output
  genomic coordinates also)
  ```

3.4 Error Messages

Please note that you may run into the following error messages while using SNVGet.

- The "M44APHC" is not recognized and therefore omitted.
  This error occurs when there is an unrecognized feature name in the Features.list file. The unrecognised feature will be ignored.

- The "MGAPHC" feature is repeated in the Features list. It has already been added to the feature set.
  This error occurs when there is a duplicated feature in the Features.list file. The duplicated feature will be ignored.

- "NP_689637 *1136Y" is not a properly formatted mutation.
  This error occurs when a row in the .tmps file is not properly formatted, or contains stop codons. This row will be ignored when retrieving feature values.

- Position/Wildtype 1126/A not found in NP_000236.
  This error occurs when the wild type amino acid for the protein does not match that stored in the database. This could be due to differences in protein version. This mutation will be ignored.

- Transcript ENST00000360484 not found in database.
  This error occurs when the database does not contain the protein sequence in question. All mutations involving that protein sequence will be ignored.

- Warning: Refseq transcript version number (NM_032173.3) does not match Refseq version in database (NM_032173.2).
  The version number of the transcript used does not match the version number of the transcript in the current version of the SNVBox database. Features will be returned from the SNVBox database for the version stored in the database.

- Sequencing variant 36915 chr22:22288398 C>T maps to NM_014634.3 L185L. Only missense variants will be evaluated by CHASM.
  The specified genomic coordinate did not map to a missense mutation. The mutation will be ignored.

- Sequencing variant chr22:20800834 A>G did not map to a codon.
  The genomic coordinate did not map to a codon. The mutation will be ignored.

- Reference base specified for sequencing variant 1 chr22:25115448 C>G does not match reference base at that coordinate in hg19.
  The reference base specified does not match the base at that position in GRCh37/hg19. The mutation will be ignored.

3.5 Available Features

The following features are currently available in SNVBox. To use each feature, simply specify the ID in the Features.list file.

ID	Feature	Description
AACharge	Net residue charge change	Change in formal charge resulting from the mutation. Histidine is assumed protonated (formal charge of +1) (Wildtype - Mutant)
AAVolume	Net residue volume change	Change in residue colume resulting from the mutation. (Wildtype - Mutant)
AAHydrophobicity	Net residue hydrophobicity change	Change in hydrophobicity resulting from the substitution. (Wildtype - Mutant)
AAGrantham	Grantham Score	Grantham substitution score for the wild type to mutant transition.
AAPolarity	Change in Polarity	Change in residue polarity due to the wildtype to mutant transition.
AAEx	Ex substitution score	Amino acid substitution score from the EX matrix.
AAPAM250	PAM250 substitution score	Amino acid substitution score from the PAM250 matrix.

AABLOSUM	BLOSUM 62 substitution score	Amino acid substitution score from the BLOSUM 62 matrix.
AAMJ	MJ Substitution score	Amino acid substitution score from the Miyazawa-Jernigan contact energy matrix.
AAHGMD2003	HGMD2003 mutation count	Number of times that the wild type to mutant substitution occursin the Human Gene Mutation Database, 2003 version.
AAVB	VB mutation score	Amino acid substitution score from the VB (Venkatarajan and Braun) matrix.
AATransition	Amino Acid Transition probabilities	Frequency of left to right transition between two neighboring amino acids based on all Uniprot Human proteins
AACOSMIC	Frequency of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database	Frequency in natural log that missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) is seen in COSMIC. These frequencies were calculated during the week of August 14, 2008, using COSMIC release 38.
AACOSMICvsSWISSPROT	Count of missense change type in the Catalog of SomaticA Mutations in Cancer (COSMIC) database divided by count in SWISSPROT database	Frequency in natural log that missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) is seen in COSMIC. These frequencies were calculated during the week of August 14, 2008, using COSMIC release 38 normalized by the occurrences of the wild type residue in human proteins found in UniProtKB.
AACOSMICvsHapMap	Count of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database divided by count in HapMap.	Frequency in natural log that missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) is seen in COSMIC. These frequencies were calculated during the week of August 14, 2008, using COSMIC release 38 normalized by the number of times the change type is observed in the HapMap SNPs database.
AAHapMap	HAPMAP Amino Acid substitution counts	Frequency in natural log the change type from Wildtype to Mutant AA that is observed in the HapMap SNPs database.
ExonConservation	46-way exon conservation	The 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. We retrieve the scores for each region that overlaps the exon in which the base substitution occurred and calculated a weighted average of the conservation scores where the weight is the number of bases with a particular

		score.
ExonSnpDensity	SNP Density	The number of SNPs in the exon where the mutation is located divided by the length of the exon.
ExonHapMapSnpDensity	HapMap verified SNP Density	The number of HapMap verified SNPs in the exon where the mutation is located divided by the length of the exon.
MGAPHC	Multiz-46-way Alignment Positional Conservation	This feature is calculated based on the degree of conservation of the residue estimated from a column in the Multiz-46-way alignment using the UCSC Human Genome Browser.
MGARelEntropy	Multiz-46-way Alignment Relative Entropy	Kullback-Leibler Distance 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.
MGAEntropy	Multiz-46-way Alignment Entropy	The Shannon entropy calculated for the column of the Multiz-46- way alignment, corresponding to the location of the mutation.
HMMPHC	Positional Hidden Markov Model (HMM) conservation score	This feature is calculated based on the degree of (HMM) conservation score conservation of the residue estimated from a multiple sequence alignment built with SAM-T2K software (29), using the protein in which the mutation occurred as the seed sequence (30). The SAM-T2K alignments are large, superfamily- level alignments that include distantly related homologs (as well as close homologs and orthologs) of the protein of interest.
HMMRelEntropy	Relative entropy of HMM alignments	Kullback-Leibler Distance 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.
HMMEntropy	Entropy of HMM alignment	The Shannon entropy calculated for the column of the SAM-T2K multiple sequence alignment, corresponding to the location of the mutation.
PredStabilityL PredStabilityM PredStabilityH	Predicted contribution to protein stability	These features consist of the probability that the wild stability type residue contributes to overall protein stability in a manner that is highly stabilizing, average or destabilizing, as predicted

		by a neural network trained with Predict-2nd software on a set of 1763 proteins with less than 30% homology. Stability estimates for the neural network training data were calculated using the FoldX force field.
PredSSE PredSSH PredSSC	Predicted secondary structure	These features consist of the probability that the secondary structure of the region in which the wild type residue exists is helix, loop or strand as predicted by a neural net trained with Predict-2nd software on a set of 1763 proteins with crystal structures and with less than 30% homology.
PredRSAB PredRSAI PredRSAE	Predicted residue solvent accessibility	These features consist of the probability of the wild type accessibility residue being buried, intermediate or exposed as predicted by a neural network trained with Predict-2nd software on a set of 1763 proteins with high- resolution X-ray crystal structures sharing less than 30% homology.
PredBFactorS PredBFactorM PredBFactorF	Predicted B-factor	These features consist of the probability that the wild type residue backbone is stiff, intermediate or flexible as predicted by a neural network trained with Predict-2nd software (29) on a set of 1763 proteins with less than 30% homology. Flexibilities for the neural net training data were estimated based on normalized temperature factors, computed using the method of (38) from the X-ray crystal structure files.
RegCompP RegCompC RegCompG RegCompDE RegCompQ RegCompH RegCompKR RegCompWYF RegCompILVM RegCompEntropy RegCompNormEntropy	Regional AA composition	The percentage of amino acids in a 15 residue window surrounding the mutation that fall into one of the following categories (P, C, G, DE, Q, H, KR, WYF, ILVM).
AATripletFirstProbWild AATripletSecondProbWild AATripletThirdProbWild	Probability of seeing the wild type residue in that position of an amino acid triple.	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB.
AATripletFirstProbMut, AATripletSecondProbMut,	Probability of seeing the mutant residue in that position of an amino acid	Calculated by joint frequencies of amino acid triples in human proteins found in UniProtKB.

AATripletThirdProbMut	triple.	
AATripletFirstDiffProb AATripletSecondDiffProb AATripletThirdDiffProb	Difference of the probability of seeing the wild type residue in that triplet position as compared with the mutant residue.	Based on the values calculated in AATripletPositionProbMut/AATripletPositionProbWild features. (Wildtype - Mutant)
UniprotBINDING UniprotACTSITE UniprotSITE UniprotLIPID UniprotMETAL UniprotCARBOHYD UniprotDNABIND UniprotNPBIND UniprotCABIND UniprotDISULFID UniprotSECYS UniprotMODRES UniprotPROPEP UniprotSIGNAL UniprotMUTAGEN UniprotTRANSMEM UniprotCOMPBIAS UniprotREP UniprotMOTIF UniprotZNFINGER UniprotREGIONS UniprotDOM_PPI UniprotDOM_RNABD UniprotDOM_TF UniprotDOM_LOC UniprotDOM_MMBRBD UniprotDOM_Chrom UniprotDOM_PostModRec UniprotDOM_PostModEnz	Uniprot Annotations (fingerprints)	These features give annotations, curated from the literature, of general binding sites, general active sites, lipid, metal, carbohydrate, DNA, phosphate and calcium binding sites, disulfides, modified residues, propeptide residues, signal peptide residues, known mutagenic sites, transmembrane regions, compositionally biased regions, repeat regions, known motifs, and zinc fingers. The integer 1 indicates that a feature is present and the integer 0 indicates that it is absent at a mutated position. The last 8 features are extracted from the DOMAIN annotation as follows: UniprotDOM_PPI :Protein- protein interaction or oligomerization UniprotDOM_RNABD:mRNA Binding UniprotDOM_TF: Transcription factor related UniprotDOM_LOC: Transport and localization related domain (localization signals, etc.) UniprotDOM_MMBRBD: Membrane binding/interacting (phosphoinositide binding, transmembrane, etc.) UniprotDOM_Chrom: Chromatin structural remodeling related domains (could be indirectly via interaction with histones, HATs ao HDACs) UniprotDOM_PostModRec: Domains that recognize or interact with post-tranlsational mod sites UniprotDOM_PostModEnz: Enzymatically active domains resulting in post-translational modification such as phosphorylation, glycosylation, amidation, ubiquitination, etc.