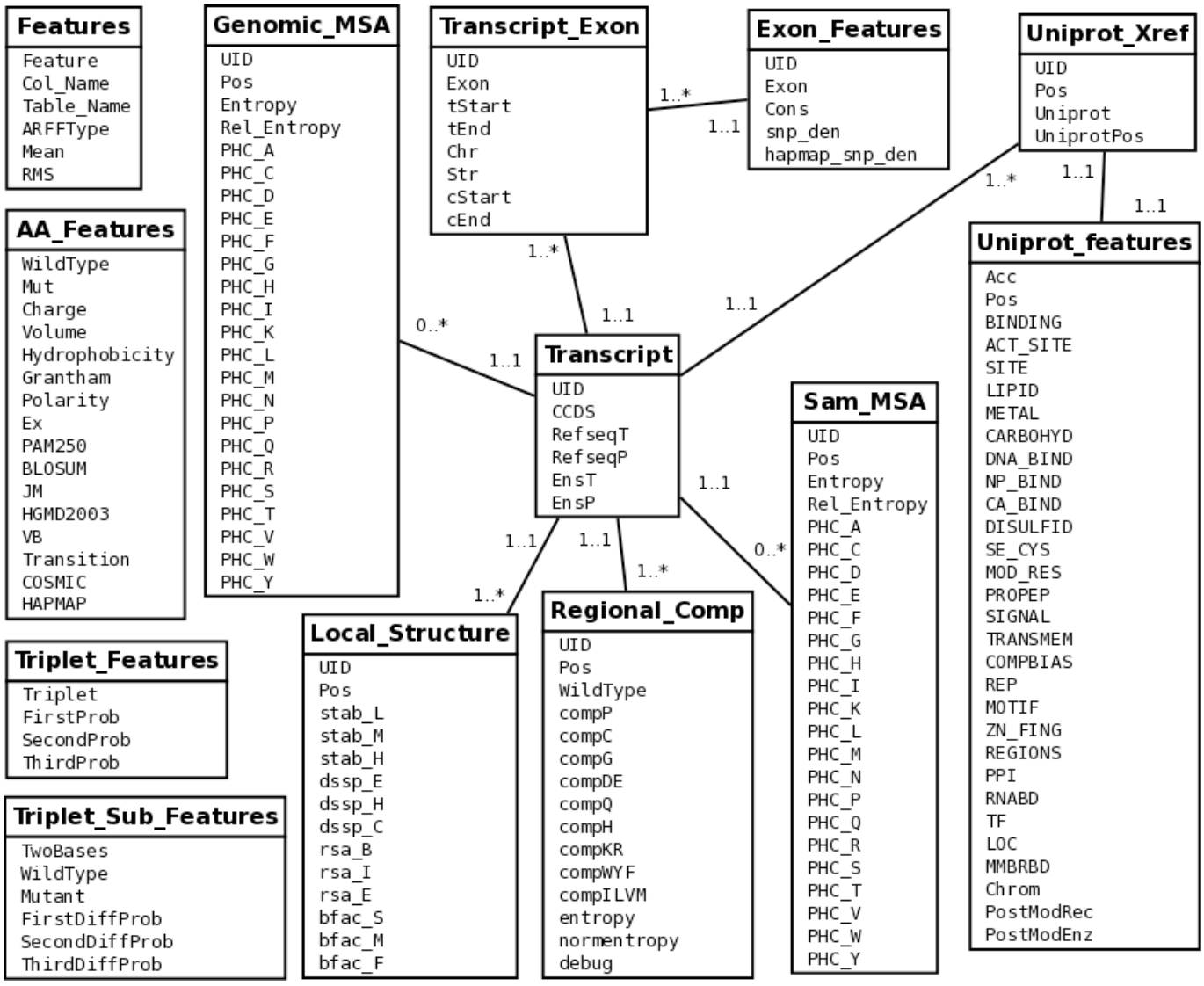
SNVBox

SNVBox is a MySQL database of 85 predictive features relevant to the biological impact of a SNV. The features have been pre-computed for each codon in all protein-coding exons of annotated human mRNA transcripts in the Refseq, CCDS and Ensembl databases.

1.1 Database Schema



1.2 Table Descriptions

Default features used by CHASM BuildClassifier highlighted in red.

1.2.1 AA_Features

A look-up table used to store amino acid substitution features. All changes are computed as (Reference amino acid residue value – Mutation amino acid residue value).

Column Name	Feature ID	Feature Name	Description

Wildtype	N/A	N/A	Reference amino acid residue.
Mut	N/A	N/A	Mutation amino acid residue.
Charge	AACharge	Net residue charge change	The change in formal charge resulting from replacing the reference amino acid residue with the mutation. Histidine is assumed protonated (formal charge of +1)
Volume	AAVolume	Net residue volume change	The change in residue volume resulting from the replacement (in units of cubic Angstroms) [1]
Hydrophobicity	AAHydrophobicity	Net residue hydrophobicity change	The change in hydrophobicity resulting from the substitution. [2]
Grantham	AAGrantham	Grantham Score	The Grantham distance from reference to mutation amino acid residue. [3]
Polarity	AAPolarity	Change in Polarity	Polarity change from reference to mutation amino acid residue [3]
Ex	AAEx	Ex substitution score	Amino acid substitution score from the EX matrix. [4]
PAM250	AAPAM250	PAM250 substitution score	Amino acid substitution score from the PAM250 matrix. [5]
BLOSUM	AABLOSUM	BLOSUM 62 substitution score	Amino acid substitution score from the BLOSUM 62 matrix [6].
JM	AAMJ	MJ Substitution score	Amino acid substitution score from the Miyazawa-Jernigan contact energy matrix [4, 7].
HGMD2003	AAHGMD2003	HGMD2003 mutation count	Number of times that the reference to mutation substitution occurs in the Human Gene Mutation Database, 2003 version [8].
VB	AAVB	VB mutation score	Amino acid substitution score from the VB (Venkatarajan and Braun) matrix [9].
Transition	AATransition	Amino Acid Transition probabilities	Frequency of transition between two neighboring amino acids based on all human proteins in SwissProt/TrEMBL [10]
COSMIC	AACOSMIC	Frequency of missense change type in the COSMIC database	Ln(frequency) of missense change type (amino acid type X to amino acid type Y, e.g. ALANINE to GLYCINE) in COSMIC (release 38) [11].
HAPMAP	AAHapMap	HAPMAP Amino Acid substitution counts	Ln(frequency) of missense change type in HapMap validated SNPs in dbSNP Build 129 [12, 13].

The following features can be derived from other columns in the table:

Feature ID	Feature Name	Description
AACOSMICvsSWISSPROT	Count of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database divided by count in SWISSPROT database	Ln(frequency) of missense change in COSMIC (release 38) [11] normalized by the frequency of reference amino acid residue in human proteins in SwissProt/TrEMBL [10].
AACOSMICvsHapMap	Count of missense change type in the Catalog of Somatic Mutations in Cancer (COSMIC) database divided by count in HapMap.	Ln(frequency) of missense change in COSMIC (release 38) [11] normalized by the number of times the change type was observed in HapMap validated SNPs in dbSNP Build 129 [12, 13].

1.2.2 Exon_Features

Features based on human variation in genomic DNA and vertebrate evolutionary conservation in human exons.

Column Name	Feature ID	Feature Name	Description
UID	N/A	N/A	Protein sequence UID
Exon	N/A	N/A	ID of the exon (Arbitrarily assigned by the program in order of occurrence in the protein)
Cons	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) [14]. 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.
snp_den	ExonSnpDensity	SNP Density	The number of SNPs in the exon where the mutation is located

			divided by the length of the exon.
hapmap_snp_den	ExonHapMapSnpDensity	HapMap verified SNP Density	The number of HapMap verified SNPs (dbSNP build 131) in the exon where the mutation is located divided by the length of the exon.

1.2.3. Features

Meta-data table used internally by SNVBox to keep track of the feature IDs of each feature and the corresponding MySQL table and Column Name under which the feature values are stored.

Column	Description
Feature	The common name of the feature stored in SNVBox. This is the name to specify to SNVget.
Col_Name	The column name of the feature as stored in the MySQL table.
Table_Name	The MySQL table in which this feature is stored in the database.
ARFF_Type	The quantity type to be specified in ARFF file format.
Mean	The average value of the column used for feature standardization.
RMS	The root-mean-square value of the column used for feature standardization.

1.2.4 Genomic_MSA

Features calculated from columns in a protein-translated version of UCSC Genome Browser's 46-way genomic vertebrate alignments [14] (computed with BLAST Multiz) [15].

Column	Feature ID	Feature Name	Description
UID	N/A	N/A	Protein sequence UID
Pos	N/A	N/A	Amino Acid position
Entropy	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.
Rel_Entropy	MGARelEntropy	Multiz-46-way Alignment Relative Entropy	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.
PHC_A	MGAPHC	Multiz-46-way	Calculated based on the degree of

		Alignment Positional Conservation	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. Feature value if the mutation is alanine.
PHC_C	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is cysteine.
PHC_D	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is aspartic acid
PHC_E	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is glutamic acid
PHC_F	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is phenylalanine
PHC_G	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is glycine
PHC_H	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is histidine
PHC_I	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is isoleucine
PHC_K	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is lysine
PHC_L	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is leucine
PHC_M	MGAPHC	Multiz-46-way	Feature value if the mutation is

		Alignment Positional Conservation	methionine
PHC_N	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is aspargine
PHC_P	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is proline
PHC_Q	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is glutamine
PHC_R	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is arginine
PHC_S	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is serine
PHC_T	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is threonine
PHC_V	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is valine
PHC_W	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is tryptophan
PHC_Y	MGAPHC	Multiz-46-way Alignment Positional Conservation	Feature value if the mutation is tyrosine

1.2.5. Local_Structure

Predicted properties of protein local structure made by predict-2nd neural network software [16]. The networks were trained using a set of 1763 high-quality protein X-ray crystal structures from divergent

proteins (all sharing less than 30% homology with the others).

Column	Feature ID	Feature Name	Description
UID	N/A	N/A	Protein sequence UID
Pos	N/A	N/A	Amino Acid position in the amino acid in the position.
stab_L	PredStabilityL	Low 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, Stability estimates for the neural network training data were calculated using the FoldX force field [17].
stab_M	PredStabilityM	Medium predicted contribution to protein stability	"
stab_H	PredStabilityH	High predicted contribution to protein stability	"
dssp_E	PredSSE	Predicted secondary structure - strand	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.
dssp_H	PredSSH	Predicted secondary structure - helix	"
dssp_C	PredSSC	Predicted secondary structure - coil	"
rsa_B	PredRSAB	Predicted residue solvent accessibility - Buried	These features consist of the probability of the wild type accessibility residue being buried, intermediate or exposed.
rsa_I	PredRSAI	Predicted residue solvent accessibility - Intermediate	"

rsa_E	PredRSAE	Predicted residue solvent accessibility - Exposed	"
bfac_S	PredBFactorS	High Predicted Bfactor	These features consist of the probability that the wild type residue backbone is stiff, intermediate or flexible.
bfac_M	PredBFactorM	Medium Predicted Bfactor	"
bfac_F	PredBFactorF	Low Predicted Bfactor	"

1.2.6 Regional_Comp

Features based on regional amino acid composition in a 15-amino-acid-residue window centered on the reference/mutation amino acid position.

Column	Feature ID	Feature Name	Description
UID	N/A	N/A	Protein sequence UID
Pos	N/A	N/A	Amino acid position
WildType	N/A	N/A	Reference amino acid of the protein at that position
compP	RegCompP	Regional AA composition	Proportion of Prolines around position.
compC	RegCompC	Regional AA composition	Proportion of Cysteines around position.
compG	RegCompG	Regional AA composition	Proportion of Glycines around position.
compDE	RegCompDE	Regional AA composition	Proportion of Aspartic and Glutamic Acids around position.
compQ	RegCompQ	Regional AA composition	Proportion of Glutamines around position.
compH	RegCompH	Regional AA composition	Proportion of Histidines around position.
compKR	RegCompKR	Regional AA composition	Proportion of Lysines and Arginines around position.
compWYF	RegCompWYF	Regional AA composition	Proportion of Tryptophans, Tyrosines, and Phenylalanines

			around position.
compILVM	RegCompILVM	Regional AA composition	Proportion of Isoleucines, Leucines, Valines, and Methionines around position.
entropy	RegCompEntropy	Regional AA composition	Shannon entropy of amino acid residues around position.
normentropy	RegCompNormEntropy	Regional AA composition	Shannon entropy of amino acid residues around position normalized by the number of different amino acids within the window.
debug	N/A	N/A	Internal flag used for debugging by the CHASM build pipeline.

1.2.7 Sam_MSA

Features calculated from a SAM-t2k hidden Markov model multiple sequence alignment [18] of diverse homologous proteins.

Column	Feature ID	Feature Name	Description
UID	N/A	N/A	Protein sequence UID.
Pos	N/A	N/A	Amino Acid position.
Entropy	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.
Rel_Entropy	HMMRelEntropy	Relative entropy of HMM alignments	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.
PHC_A	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	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. Feature value if the mutation is alanine.

PHC_C	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is cysteine.
PHC_D	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is aspartic acid
PHC_E	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is glutamic acid
PHC_F	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is phenylalanine
PHC_G	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is glycine
PHC_H	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is histidine
PHC_I	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is isoleucine
PHC_K	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is lysine
PHC_L	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is leucine

PHC_M	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is methionine
PHC_N	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is aspargine
PHC_P	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is proline
PHC_Q	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is glutamine
PHC_R	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is arginine
PHC_S	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is serine
PHC_T	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is threonine
PHC_V	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is valine
PHC_W	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is tryptophan

PHC_Y	HMMPHC	Positional Hidden Markov Model (HMM) conservation score	Feature value if the mutation is tyrosine

1.2.8 Transcript

Accession identifiers for mRNA in CCDS (and for mRNA translated protein) in RefSeq and Ensembl.

Column Name	Description
UID	Protein sequence UID
CCDS	Consensus CDS mRNA accession
RefseqT	Refseq mRNA accession
RefseqP	Refseq protein accession
EnsT	Ensembl mRNA accession
EnsP	Ensembl protein accession

1.2.9 Transcript_Exon

Exon boundaries in the mRNA transcript associated with each protein sequence.

Column Name	Description
UID	Protein sequence UID
Exon	Identifier of the exon within each mRNA transcript
tStart	Starting position of the exon within the mRNA transcript
tEnd	Ending position of the exon within the mRNA transcript
Chr	Chromosome containing the exon
Str	Direction of transcription
cStart	Starting position of the exon within the chromosome
cEnd	Ending position of the exon within the chromosome

1.2.10 Triplet_Features

Features based on frequencies of contiguous amino-acid residue triplets in human protein sequences in SwissProt/TrEMBL.

Column Name	Feature ID	Feature Name	Description

Triplet	N/A	N/A	All possible combinations of 3 adjacent amino acids.
FirstProb	AATripletFirstProbWild, AATripletFirstProbMut	First position probability	Probability of seeing the amino acid in position 1 of a triplet
SecondProb	AATripletSecondProbWild, AATripletSecondProbMut	Second position probability	Probability of seeing the amino acid in position 2 of a triplet
ThirdProb	AATripletThirdProbWild, AATripletThirdProbMut	Third position probability	Probability of seeing the amino acid in position 3 of a triplet

The following features can be derived from other columns in the table:

Feature ID	**Feature Name**	**Description**
AATripletFirstDiffProb	First position probability change	Difference in probability of occurrence of reference and mutation amino acid residue in the 1st position.
AATripletSecondDiffProb	Second position probability change	Difference in probability of occurrence of reference and mutation amino acid residue in the 2nd position.
AATripletThirdDiffProb	Third position probability change	Difference in probability of occurrence of reference and mutation amino acid residue in the 3rd position.

1.2.11 Uniprot_Xref

Cross-reference of accession ID and amino acid residue positions of translated mRNA transcript with canonical SwissProt protein sequence.

Column Name	**Description**
UID	Protein sequence UID
Pos	Position of the protein
Uniprot	Best canonical SwissProt protein match using BLAST
UniprotPos	Alignment of the protein position to SwissProt amino acid position using BLAST

1.2.12 Uniprot_features

Uniprot feature annotations of human protein sequences.

Column Name	Feature ID	Feature Name	Description
Acc	N/A	N/A	Uniprot accession
Pos	N/A	N/A	Position in the uniprot sequence
BINDING	UniprotBINDING	Uniprot Annotations	Binding sites. The integer 1 indicates that a feature is present and the integer 0 indicates that it is absent.
ACT_SITE	UniprotACTSITE	Uniprot Annotations	Sites involved in enzymatic activity
SITE	UniprotSITE	Uniprot Annotations	An interesting amino acid site in the protein sequence
LIPID	UniprotLIPID	Uniprot Annotations	Lipid binding site
METAL	UniprotMETAL	Uniprot Annotations	Metal binding site
CARBOHYD	UniprotCARBOHYD	Uniprot Annotations	Carbohydrate binding site
DNA_BIND	UniprotDNABIND	Uniprot Annotations	DNA binding site
NP_BIND	UniprotNPBIND	Uniprot Annotations	Nucleotide phosphate-binding region
CA_BIND	UniprotCABIND	Uniprot Annotations	Calcium binding site
DISULFID	UniprotDISULFID	Uniprot Annotations	Site of disulfide bond
SE_CYS	UniprotSECYS	Uniprot Annotations	Site of a selenocystein
MOD_RES	UniprotMODRES	Uniprot Annotations	Site of modified residue
PROPEP	UniprotPROPEP	Uniprot Annotations	Site in the propeptide (cleaved in mature protein)
SIGNAL	UniprotSIGNAL	Uniprot Annotations	Site of localization signal (protein targeted to secretory pathway or periplasm)
TRANSMEM	UniprotTRANSMEM	Uniprot Annotations	Transmembrane region

COMPBIAS	UniprotCOMPBIAS	Uniprot Annotations	Compositionally biased region
REP	UniprotREP	Uniprot Annotations	Repeat region
MOTIF	UniprotMOTIF	Uniprot Annotations	Site of known functional motif
ZN_FING	UniprotZNFINGER	Uniprot Annotations	Site in a zinc finger
REGIONS	UniprotREGIONS	Uniprot Annotations	Region of interest in the protein sequence
PPI	UniprotDOM_PPI	Uniprot Annotations	Site in a protein-protein interaction domain
RNABD	UniprotDOM_RNABD	Uniprot Annotations	Site in an RNA binding domain
TF	UniprotDOM_TF	Uniprot Annotations	Site in a transcription factor doman
LOC	UniprotDOM_LOC	Uniprot Annotations	Site in a domain that determines correct cellular localization of a protein
MMBRBD	UniprotDOM_MMBRBD	Uniprot Annotations	Site in a domain that binds to the cell membrane
Chrom	UniprotDOM_Chrom	Uniprot Annotations	Site in a domain involved in chromatin structure remodeling
PostModRec	UniprotDOM_PostModRec	Uniprot Annotations	Site in a domain that recognizes a post-translationally modified residue
PostModEnz	UniprotDOM_PostModEnz	Uniprot Annotations	Site in an enzymatic domain responsible for any kind of post-translational modification

1.3 SNVBox Utilities

The SnvGet program is provided for users to query the SNVBox database to quickly retrieve features describing a mutation of interest, by specifying an accession number for mRNA transcript or translated protein (RefSeq, CCDS, or Ensembl), codon position, reference and mutation amino acid residues, and a list of requested features. Results are output in ARFF file format, allowing a user to run custom machine learning algorithms or standard algorithms available in packages such as WEKA, WAFFLES and R. Users can also query the database directly, using the provided schema and table descriptions (Supp Information 1.1, 1.2).

References

1. Zamyatnin, A.A., *Protein volume in solution.* Prog Biophys Mol Biol, 1972. **24**: p. 107-23.
2. Engelman, D.M., T.A. Steitz, and A. Goldman, *Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins.* Annu Rev Biophys Biophys Chem, 1986. **15**: p. 321-53.
3. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. **185**(4154): p. 862-864.
4. Yampolsky, L.Y. and A. Stoltzfus, *Untangling the effects of codon mutation and amino acid exchangeability.* Pac Symp Biocomput, 2005: p. 433-44.
5. Schwartz, R.M. and M.O. Dayhoff, *IMPROVED SCORING MATRIX FOR IDENTIFYING EVOLUTIONARY RELATEDNESS AMONG PROTEINS.* Biophys J, 1978. **21**(3): p. A198-A198.
6. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-10919.
7. Miyazawa, S. and R.L. Jernigan, *Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.* Macromolecules, 1985. **18**(3): p. 534-552.
8. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-581.
9. Venkatarajan, M.S. and W. Braun, *New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties.* Journal of Molecular Modeling, 2001. **7**(12): p. 445-453.
10. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information*, 2006.
11. Forbes, S., et al., *Cosmic 2005.* Br J Cancer, 2006. **94**(2): p. 318-22.
12. Schmidt, C.W., *HapMap: building a database with blocks.* EHP Toxicogenomics, 2003. **111**(1T): p. A16.
13. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
14. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996--1006.
15. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
16. Katzman, S., et al., *PREDICT-2ND: a tool for generalized protein local structure prediction.* Bioinformatics, 2008. **24(21)**: p. 2453-9.
17. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
18. Karplus, K., et al., *What is the value added by human intervention in protein structure prediction?* Proteins, 2001. **Suppl 5**: p. 86-91.