**Authors**

Qiong Wei

Roland L. Dunbrack, Jr.

Institute for Cancer Research

Fox Chase Cancer Center

333 Cottman Avenue

Philadelphia PA 19111

USA

Contact: Roland.Dunbrack@fccc.edu

Phone: 215 728 2434

Fax: 215 728 2412

***Data sets***

We created a dataset, *HumanDisease*, from the SwissVar database1 (release of 2012\_03) by removing unclassified variants, variants in very long proteins (sequences of more than 3000 amino acids), and redundant variants. We obtained non-human primate sequences from UniprotKB2 and used PSI-BLAST3,4 to identify likely primate orthologues of human proteins in the SwissVar data sets using a sequence identity cutoff of 90% between the human and primate sequences. To avoid misaligned amino acids near deletions, perhaps caused by missing exons, we selected primate sequence differences without insertions or deletions within 10 amino acids on either side of the position of amino acid differences in the PSI-BLAST alignments, and compiled them into a data set of human/primate sequence differences, *PrimateMut*.

To define the structural dataset, we mapped the human mutation sites in the *HumanDisease* and *PrimateMut* data sets to known structures of human proteins in the PDB using SIFTS, which provides Uniprot sequence identifiers and sequence positions for residues in the PDB. We constructed two structural data sets, *HumanDiseaseStr* and *PrimateMutStr*, which contain mutation sites available in known structures.

We compiled a set of 6737 human disease mutations with structures in the PDB (i.e. structures of the human protein, not a homologue thereof), while the other 14979 human disease mutations do not have native structures in the PDB. We also found that the positions of 5671 human/primate sequence differences are available in structures of the human proteins in the PDB, while 28148 primate sequence differences do not have native human structures in PDB. We combined human disease mutations with structure and primate sequence differences with structures together as our structure-based training data set, and we combined human disease mutations without structures and primate sequence differences without structures together as our sequence-based data set. For our two data sets, we calculate structure-based, sequence-based, and disorder/order features to describe the mutations.

***Structure-based features***

For wildtype position of each mutation, we used the program Naccess5 to calculate the percent accessible surface area in all available biological assemblies and monomers containing the mutation site. For biological assemblies, we downloaded the XML file from PISA which contains symmetry operators for building the coordinates of these assemblies. For each mutation, we calculated the accessible surface area in all available biological assemblies and monomers, and then selected the biological assembly with the minimal accessible surface to get the biological accessible surface area and monomer surface area.

***Sequence-based features***

We used PSI-BLAST to search human protein sequences against the database UniRef1006 for two rounds with an E-value cutoff of 0.002 for inclusion in the Position-Specific Scoring Matrix (PSSM) score for the mutations. We obtained the PSSM score of the wildtype residues and the PSSM score of the mutant residues from the PSI-BLAST matrix output. For each query, we selected homologues from the PSI-BLAST output with sequence identity greater than 20% and input these proteins to the program BLASTCLUST3 to cluster these sequences at a sequence identity threshold of 35%. A multiple sequence alignment of the sequences in the cluster containing the target human protein was created with the program Muscle7,8. The multiple sequence alignment was input to the program AL2CO9 to calculate the conservation score for each wildtype residue of human proteins.

Excluding PSSM score and conservation score, we also calculated the side chain volume change from wildtype residue to mutant residue and the hydrophobic change from wildtype residue to mutant residue.

We ran three disorder prediction programs VSL210, IUpred11 and Espritz12 to perform disorder prediction for all human proteins in *HumanDisease* and *PrimateMut*. We used the outputs of those three programs as the features for each mutation in *HumanDisease* and *PrimateMut* data sets.

***Training and Testing Data Features***

Ultimately, the sequence-based data set contains seven features (difference between wildtype PSSM and mutant PSSM; conservation score; side chain volume change; hydrophobic change; IUPred output value, Espritz output value and VSL output value), and the structure-based data set contains nine features (accessible surface area of biological assembly and protein monomer, and the other same seven features as the sequence-based data set).

***Testing Data***

For each mutation in testing data, we try to find structure by using SIFT to get actual structure in PDB and our homology modeling program BioAssemblyModeler (BAM) (Shapovalov et al., submitted to *Bioinformatics*) to get predicted biological assembly structure so that we can calculate the accessible surface area. We separated the testing mutations into two subsets, one that contains mutations that have an actual structure or predicted structure and the other which contains the mutations which do not have structure information, and then use different SVM models to predict the phenotype for each subset.

***Support Vector Machine***

We randomly balanced our two training datasets so that we obtain reliable performance (Wei and Dunbrack, *PLOSOne*, in press). The structure-based training set contains 11342 mutations, and the sequence-based training set in contains 29958 mutations. The Support vector machine program, SVMLight, was used to train a structure-based model and sequence-based model based on those two balanced training data sets.

The expression



where  is the SVM output of mutation *i*, was used to translate the SVM output to probability that a mutation was predicted as positive and 10-fold cross-validation was used to get optimal parameters A and B. Once we get a testing data set without known phenotype, we calculate the features for them, and separate them into two subsets according the availability of structure, and then select one of the two SVM models to predict the phenotype.

***Special Cases***

Synonymous mutations are the evolutionary substitutions of one base for another in an exon of a gene coding for a protein, such that the produced amino acid sequence is not modified. When a synonymous mutation occurs, the change is often assumed to be neutral, so in our program, if the mutation is synonymous, we arbitrarily assign the mutation a 5% probability of being deleterious. Of course some synonymous mutations may effect transcription, splicing, or translation, but we do not have a method for accounting for this.

Nonsense mutations are changes in DNA sequence that introduce a premature stop codon, causing any resulting protein to be abnormally shortened. This often causes a loss of function in the protein, as critical parts of the amino acid chain are no longer created. So our program predicts stop codon mutations as deleterious with 95% probability. Deletions were handled similarly, except in one case where the deletion was in a disordered region and the mutation of each deleted residue to Gly was predicted to be neutral (probably because the residues did not have high conservation). We treated this deletion as neutral.

**References**

1. Mottaz A, David FP, Veuthey AL, Yip YL. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics 2010;26(6):851-852.

2. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford);2011:bar009.

3. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of database programs. Nucleic Acids Research 1997;25:3389-3402.

4. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK. Protein database searches using compositionally adjusted substitution matrices. Febs J 2005;272(20):5101-5109.

5. Hubbard SJ, Thornton JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College London; 1993.

6. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32(Database issue):D115-119.

7. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5(1):113.

8. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32(5):1792-1797.

9. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17(8):700-712.

10. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.

11. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21(16):3433-3434.

12. Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012;28(4):503-509.