Supplemental Material: Annu.Rev. Genom. Human. Genet. 2006. 7: 61-80
doi: 10.1146/annurev.genom.7.080505.115630
Predicting the Effects of Amino Acid Substitutions on Protein Function
Ng and Henikoff

Supplemental Table 1 A survey of AAS prediction methods and their observations.

Groups with key observations have been highlighted in bold.

Group	Observations
Wang and Moult (2001) (77)	●Observed that disease mutations can be distinguished from neutral SNPs using structure. ●Modeled amino acid change on to structure and observed that 83% of disease SNPs affect protein stability; 5% affect ligand binding.
Sunyaev et al. (2000, 2001, 2002) (53, 68, 69)	●Prediction method incorporates sequence conservation, structure, and annotation from Swiss-Prot database for prediction. ●~20% of nsSNPs estimated to be damaging to protein function. To resolve the large number of damaging nsSNPs with the fecundity of inbred marriages, damaging nsSNPs must have mild effects.
Chasman and Adams (2001) (9)	●Accessibility, *B*-factor and sequence conservation are the most useful for prediction according to ANOVA, PCA and correlation analysis. Implemented a prediction method based on the three features. ●A protein structure that has >= 60% sequence identity to the input protein gives the best performance, with no increase in performance as sequence identity increases.
Ng and Henikoff (2001,2002, 2003) (45--47)	●Prediction method uses sequence conservation and position-specific scoring matrices with Dirichlet priors to model the allowed amino acids at a position. ●Using an AAS prediction method has better accuracy than an amino acid substitution scoring matrix. ●The fraction of nsSNPs predicted to be damaging is comparable to the false positive error so the number of damaging nsSNPs is low and cannot be estimated.
Ferrer-Costa et al. (16--18) (2002, 2004, 2005)	●Sequence-based prediction method. Structural properties (secondary structure and accessibility) are predicted and used for prediction. A neural network provides the final prediction.
Saunders and Baker (2002) (59)	●Found that the most accurate predictions are obtained using a combination of sequence and structural features. ●Estimated C^{β} density values from ab initio structure prediction which can be beneficial when there are few sequences available (<= 3 homologues).
Terp et al. (2002) (72)	●Studied 20 biophysical parameters, 9 of which were significant for prediction. ●Score includes the likelihood that a mutation will be observed clinically. The disease-causing mutation is severe enough to be included in a disease database, but not preclinically lethal.
Mooney et al. (2003,2003) (41, 42)	●Using sequence homology, calculate negative entropy to measure conservation. ●Found degree of conservation scores for mutations are correlated with the severity of the phenotype for syndromes caused by mutations in the androgen receptor.

Group	Observations
Stitziel et al. (2003, 2004) (64, 65)	●TopoSNP uses sequence conservation and structure. Structure is used to classify if the AAS is located on a surface, in a pocket, or buried. Under this classification, only 3% of disease-associated mutations were buried. ●Using structure, 88% of disease-associated mutations are in pockets but 68% of nondisease SNPs are also in these pockets.
Krishnan and Westhead (2003) (31)	●Implemented 2 different machine-learning methods: decision trees and support vector machines. ●Methods use sequence and structure. Structural attributes (secondary structure and solvent accessibility) are predicted and homologous structure is not necessary.
Thomas et al. (2003) (73)	●Position-specific evolutionary conservation scores calculated from Hidden Markov Models in the PANTHER library. ●Some disease-causing mutations predicted to be gain-of-function rather than loss-of-function.
del sol Mesa et al. (2003) (14)	●Implemented three different methods to partition a protein family into subfamilies for better prediction of positions involved in functional specificity.
Fleming et al. (2003) (19)	●Identify conserved sites through sliding window. ●Incorporates DNA sequence to identify sites evolving under positive selection.
Santibáñez-Koref et al. (2003) (58)	●Uses sequence conservation and takes into account phylogeny. ●Altering tree structure decreases performance; altering branch lengths does not affect performance as much.
Herrgard et al. (2003) (27)	●Using sequence and structure, this prediction method focuses on mutations in the active site of an enzyme to find residues that affect catalytic ability rather than protein stability. ●Average prediction accuracy for deleterious mutations 85%, prediction for nondeleterious mutations is 81%.
Cai et al. (2004) (7)	●Bayesian network evaluating only positions residing inside a PFAM domain. Also uses structural annotation provided by SWISS-PROT.
Lau and Chasman (2004)(33)	●From a set of sequences, chooses the optimal sequence subalignment which gives as many tolerated amino acids but excludes proteins that are functionally divergent from the query protein. ●Performs better than SIFT.
Balasubramanian et al. (2005)(2)	●Uses logistic regression analysis based on sequence and structural features to obtain high prediction accuracy on G-protein coupled receptors.
Stone and Sidow (2005) (67)	●Sequence-based MAPP method, which performs better than SIFT. ●Takes a protein alignment and a tree for input. ●Quantified the benefit of using orthologues instead of paralogues in sequence alignment.

Group	Observations
Yue and Moult (2005) (82, 83)	●Two methods based on support vector machine. One is sequence-only, one is structure-only. ●Larger change in free energy is correlated with a higher fraction of disease mutations. ●SwissProt functional annotation decreases overall accuracy. ●Protein models based on structure with >= 40% sequence identity has comparable performance to using the structure of the input protein. ●Sequence-based method has better overall performance compared to structure. ●When number of sequences is < 10, false positive rate increases but false negative rate stays the same.