Human Mutation
OFFICIAL JOURNAL
HGVS HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Performance of Mutation Pathogenicity Prediction Methods on Missense Variants

Janita Thusberg,[1,2] Ayodeji Olatubosun,[1] and Mauno Vihinen[1,3]*

[1]*Institute of Biomedical Technology, F1-33014 University of Tampere, Finland;* [2]*Buck Institute for Age Research, Novato, California;* [3]*Research Center, Tampere University Hospital, Tampere, Finland*

Communicated by Christophe Béroud

Received 20 May 2010; accepted revised manuscript 7 December 2010.
Published online 22 February 2011 in Wiley Online Library (www.wiley.com/humanmutation). DOI 10.1002/humu.21445

ABSTRACT: Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans. The number of SNPs identified in the human genome is growing rapidly, but attaining experimental knowledge about the possible disease association of variants is laborious and time-consuming. Several computational methods have been developed for the classification of SNPs according to their predicted pathogenicity. In this study, we have evaluated the performance of nine widely used pathogenicity prediction methods available on the Internet. The evaluated methods were MutPred, nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP, and SNPs&GO. The methods were tested with a set of over 40,000 pathogenic and neutral variants. We also assessed whether the type of original or substituting amino acid residue, the structural class of the protein, or the structural environment of the amino acid substitution, had an effect on the prediction performance. The performances of the programs ranged from poor (MCC 0.19) to reasonably good (MCC 0.65), and the results from the programs correlated poorly. The overall best performing methods in this study were SNPs&GO and MutPred, with accuracies reaching 0.82 and 0.81, respectively.
Hum Mutat 32:358–368, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: method evaluation; bioinformatics; pathogenicity prediction; SNPs

Introduction

Most human genetic variation is represented by single nucleotide polymorphisms (SNPs), and many of them are believed to cause phenotypic differences between individuals. Owing to the application of high-throughput sequencing methods, the number of identified variants in the human genome is growing rapidly, but identifying those variations responsible for specific phenotypes is a laborious task. The ability to discriminate between pathogenic and benign variants computationally could significantly aid targeting disease-causing mutations by helping in the selection and prioritization of likely candidates from a pool of data. A subset of SNPs occur at protein coding regions in the genome, and from a medical point of view particularly interesting ones are the nonsynonymous SNPs (nsSNPs) that lead to an amino acid substitution at the protein level (referred here to as missense variants). nsSNPs may affect gene function through their effect on the structure and/or function of the encoded protein.

Prediction of the possible disease-association of missense variants is a difficult problem because an amino acid substitution can affect the biological function of a gene product in a number of ways [Thusberg and Vihinen, 2009]. An amino acid substitution may disrupt sites that are critical in protein function, such as catalytic residues or ligand-binding pockets. A missense mutation may as well lead to alterations in the structure, folding, or stability of the protein product, thereby altering or preventing the function of the protein. On the other hand, amino acid substitutions do not necessarily affect protein function. Effects of missense mutations are often the most difficult to predict while the consequences of most deletions, insertions, and nonsense mutations are rather self-evident.

Many methods have been developed for the computational prediction of the phenotypic effect of nsSNPs. Some of them are for the study of very specific mechanisms, whereas others are developed to predict whether a variation is harmful or benign. All of the variation tolerance methods evaluated in this study follow a similar procedure in which a missense variant is first labeled with properties related to the damage it may cause to the protein structure or function. The resulting feature vector is then utilised to decide whether the variant is pathogenic or not. The methods differ in the properties of the variant they take into account in the prediction, as well as in the nature and possible training of the classification method used for decision making. The nine widely used methods evaluated in this study are based on evolutionary information (Panther [Thomas et al., 2003], PhD-SNP SVM-Profile [Capriotti et al., 2006], and SIFT [Ng and Henikoff, 2001]), or a combination of protein structural and/or functional parameters and multiple sequence alignment derived information (MutPred [Li et al., 2009], nsSNPAnalyzer [Bao et al., 2005], PolyPhen [Ramensky et al., 2002], PolyPhen2 [Adzhubei et al., 2010], SNAP [Bromberg and Rost, 2007], and SNPs&GO [Calabrese et al., 2009]). The machine-learning methods utilize neural networks (NN) (SNAP), random forests (RF) (MutPred, nsSNPAnalyzer), or support vector machines (SVMs) (PhD-SNP, SNPs&GO) for classification, whereas the other methods classify variants according to empirically derived rules (PolyPhen), Bayesian methods (PolyPhen2), or mathematical operations (SIFT, Panther) (Table 1).

*Correspondence to: Mauno Vihinen, Institute of Biomedical Technology, F1-33014 University of Tampere, Finland. E-mail: mauno.vihinen@uta.fi

Contract grant sponsors: The Tampere Graduate School in Biomedicine and Biotechnology; The Sigrid Jusélius Foundation; The Academy of Finland; The Medical Research Fund of Tampere University Hospital.

Table 1. Summary of the Evaluated Methods

Method	Based on	Training set	Conservation analysis	Structural attributes	Annotations	Website
MutPred	RF	HGMD, Swiss-Prot	SIFT, Pfam, PSI-BLAST	Predicted attributes	–	http://mutpred.mutdb.org/
nsSNPAnalyzer	RF	Swiss-Prot	SIFT	Homologue mapping	–	http://snpanalyzer.uthsc.edu/
Panther	Alignment scores	–	Panther library, HMMs	–	–	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PhD-SNP	SVM	Swiss-Prot	Sequence environment, sequence profiles	–	–	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
PolyPhen	Empirical rules	–	PSIC profiles	Homologue mapping/predictions	Swiss-Prot	http://genetics.bwh.harvard.edu/pph/
PolyPhen2	Bayesian classification	Swiss-Prot, neutral pseudo-mutations	PSIC profiles	Homologue mapping/predictions	Pfam domain	http://genetics.bwh.harvard.edu/pph2/
SIFT	Alignment scores	–	MSAs	–	–	http://sift.jcvi.org/
SNAP	NN	PMD, neutral pseudo-mutations	PSIC profiles, Pfam, PSI-BLAST	Predictions	–	http://rostlab.org/services/snap/
SNPs&GO	SVM	Swiss-Prot	Sequence environment, sequence profiles, Panther	–	GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/

GO, Gene Ontology; HGMD, Human Gene Mutation Database; HMM, Hidden Markov model; NN, neural network; MSA, multiple sequence alignment; PMD, Protein Mutant Database; PSIC, position-specific independent counts; RF, random forest; SVM, support vector machine.

As mutation data and information about the genotypes of individuals accumulate, understanding the molecular level effects of variations and elucidating their possible disease association is an important research challenge [Karchin, 2009; Mooney, 2005; Ng and Henikoff, 2006; Steward et al., 2003; Thusberg and Vihinen, 2009]. Numerous locus-specific databases (LSDBs) have been established for the collection, analysis, and distribution of disease-related variation information in certain genes. Data for several genes is available, for example, in the protein knowledgebase SwissProt [Yip et al., 2004] and PhenCode [Giardine et al., 2007], which is a database that connects human variant data with phenotypic information from LSDBs with genomic data from the ENCODE project and other resources in the UCSC Genome Browser [Raney et al., 2011]. SNP information is available in dbSNP [Sherry et al., 2001], a genetic variation database. Several tools for the prediction of the phenotypic consequences of missense variants are available, but without knowledge about the quality of predictions, choosing the best method and evaluating the reliability of its outcome is impossible. We therefore performed the first comprehensive systematic evaluation of nine bioinformatics tools predicting the phenotypic effects of missense variants.

Materials and Methods

Datasets

We built a positive dataset (referred to as pathogenic dataset) of 19,335 missense mutations from the PhenCode database [Giardine et al., 2007] (downloaded in June 2009), registries in IDbases [Piirilä et al., 2006] and from 18 individual LSDBs, and a negative (neutral) dataset of 21,170 human nonsynonymous coding SNPs with an allele frequency >0.01 and chromosome sample count >49 from the dbSNP database [Sherry et al., 2001] build 131. The SNP data was filtered so that none of the dbSNP entries included in our dataset contained OMIM links to minimize the number of disease-associated SNPs in the neutral dataset. Entries annotated as "putative" or "predicted" were also left out. In addition, the neutral dataset was searched against the pathogenic dataset in order to remove possible duplicates and further minimise the probability of having false negative cases in the set. The PhenCode data was filtered so that only SNPs annotated as disease causing in the SwissProt database were taken into our pathogenic dataset. Swiss-Prot provides high-quality hand-curated information about the possible disease-relation of nsSNPs, derived from literature [Yip et al., 2008]. The complementing LSDB data was retrieved manually from each database. The pathogenic and neutral datasets contained 1,190 and 9,011 proteins, respectively, of which 445 and 1,205 were found to have three-dimensional structure coordinates in the Protein Data Bank (PDB) [Berman et al., 2000]. The datasets are available for download at our Website (http://bioinf.uta.fi).

Both datasets were run by all of the nine methods studied here. The number of results from nsSNPAnalyzer is much smaller than the original number of cases in the input data, because the program only accepts mutations in those sequences for which a homologous protein is found in the ASTRAL database [Chandonia et al., 2004]. A large number of proteins in our dataset did not match with any entry in the database, thus limiting the number of cases that could be analysed by nsSNPAnalyzer.

Two kinds of subdatasets were constructed from the original pathogenic and neutral datasets. First, a structural subdataset was compiled from the part of both datasets for which structural data was available in the PDB, to study the effect of available structure data on prediction performance. Second, for probing the effect of using Swiss-Prot-derived data as part of the pathogenic testing set, we constructed a subdataset containing only pathogenic variants not present in Swiss-Prot. The corresponding neutral dataset was compiled by randomly selecting an equal number of variants from the original neutral test set.

To test whether the differences in method performance with these subdatasets was caused by smaller testing set size, we constructed 100 sample datasets each containing 1,000 pathogenic and 1,000 neutral variants randomly picked from the original datasets, and compared the average MCCs obtained with the MCCs from the subdatasets.

The Pathogenic-or-not Pipeline (PON-P) [Thusberg and Vihinen, 2009] was used for the submission of sequences and variants into the analysis programs nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, PolyPhen2, SIFT, and SNAP. PON-P is a service that simultaneously submits the input data provided by the user to selected prediction methods. MutPred and SNPs&GO were run locally at the corresponding laboratories by the developers of the methods.

Prediction Methods

The effects of mutations and SNPs were predicted by the programs MutPred [Li et al., 2009], nsSNPAnalyzer [Bao et al., 2005], Panther [Thomas et al., 2003], PhD-SNP [Capriotti et al., 2006], PolyPhen [Ramensky et al., 2002], PolyPhen2 [Adzhubei et al., 2010], SIFT [Ng and Henikoff, 2001], SNAP [Bromberg and Rost, 2007], and SNPs&GO [Calabrese et al., 2009]. Key properties of the methods are listed in Table 1. The default parameters of all programs were applied, and only the protein sequence and missense variant were given as input information for each program, as in a normal user situation of unknown variant analysis.

MutPred

MutPred is a Random Forest-based classification method that utilizes several attributes related to protein structure, function, and evolution. MutPred utilizes the SIFT method [Ng and Henikoff, 2003] for defining the evolutionary attributes, along with PSI-BLAST, transition frequencies [Bromberg and Rost, 2007], and Pfam profiles [Finn et al., 2010]. In MutPred, structural descriptors include prediction of secondary structure and solvent accessibility by the method PHD [Rost, 1996], transmembrane helix prediction by TMHMM [Krogh et al., 2001], coiled-coil structure prediction by MARCOIL [Delorenzi and Speed, 2002], stability prediction by I-Mutant 2.0 [Capriotti et al., 2005], B-factor prediction [Radivojac et al., 2004], and disorder prediction by DisProt [Peng et al., 2006]. Function-related attributes include predictions of DNA-binding residues [Ahmad et al., 2004], catalytic residues, calmodulin-binding targets [Radivojac et al., 2006], and posttranslational modification sites [Daily et al., 2005; Iakoucheva et al., 2004; Radivojac et al., 2010]. The MutPred method estimates effects of an amino acid substitution on the set of defined properties of a protein and based on those estimates, predicts whether an amino acid substitution is likely to have phenotypic effects.

nsSNPAnalyzer

nsSNPAnalyzer is a machine-learning method that integrates multiple sequence alignment (MSA) and protein structure analysis to classify missense variants. The input protein sequence is searched against the ASTRAL database [Chandonia et al., 2004] for homologous protein structures, and extracts features of the environment of the substitution from the obtained structure, namely, the solvent accessibility, environmental polarity, and secondary structure. The SIFT method [Ng and Henikoff, 2003] is used for calculating the normalised probability of the substitution in the MSA, and the similarity and dissimilarity between the mutated, that is, original, and mutant residue is also taken into account. The program then uses a Random Forest classifier trained by a dataset prepared from the Swiss-Prot database [Yip et al., 2004] to classify the variant to be disease-associated or functionally neutral.

Panther

The Panther Evolutionary Analysis of Coding SNPs (referred simply to as Panther in this article) calculates substitution position-specific evolutionary conservation (subPSEC) scores based on alignments of evolutionarily related proteins to predict the pathogenicity. The alignments are obtained from the PANTHER library of protein families based on Hidden Markov Models (HMMs). The subPSEC score describes the amino acid probabilities, in particular, positions among evolutionarily related sequences, and the values range from 0 (neutral) to about -10 (most likely to be deleterious). The cutoff for classifying a missense variant to be pathogenic can be defined by the user, but the authors of the method advice to use a cutoff of -3 for classification [Thomas et al., 2003].

PhD-SNP

PhD-SNP is a prediction method based on single-sequence and sequence profile based support vector machines trained on Swiss-Prot variants [Yip et al., 2004]. The single-sequence SVM (SVM-Sequence) classifies the missense variant to be pathogenic or neutral based on the nature of the substitution and properties of the neighboring sequence environment. The profile-based SVM (SVM-Profile) utilizes sequence profile information taken from MSAs, and classifies the variant according to the ratio between the frequencies of the wild-type and substituted residue. A decision tree algorithm chooses which one of the two SVMs described above is to be used at each case based on the occurrence of wild-type and mutant amino acids at the given position.

PolyPhen

PolyPhen (Polymorphism Phenotyping) uses a rule-based cutoff system to classify variants. It initially characterises the input missense variant by various sequence, structure, and phylogeny based descriptors. The sequence-based characterisation includes SWALL database [Johnson and Todd, 2000] annotations for sequence features, a transmembrane predictor TMHMM [Krogh et al., 2001] and PHAT [Ng et al., 2000] transmembrane-specific matrix score for substitutions at predicted transmembrane regions, the Coils2 program [Lupas et al., 1991] for prediction of coiled coil regions, and the SignalP [Nielsen et al., 1997] program to predict signal peptide regions. Phylogenetic information is derived by constructing a profile matrix from aligned sequences by the PSIC (Position-Specific Independent Counts) software [Sunyaev et al., 1999]. The structural descriptors are obtained by mapping the missense variant onto the corresponding or similar protein and then using the DSSP program [Kabsch and Sander, 1983] for secondary structure information, solvent-accessible surface, and φ–ψ dihedral angles. In addition, PolyPhen calculates the normalized accessible surface area and changes in accessible surface propensity resulting from the amino acid substitution, change in residue side chain volume, region of the Ramachandran map, normalized B factor, and loss of a hydrogen bond according to the Hbplus program [McDonald and Thornton, 1994]. The SWALL database annotations are utilized in the structure analysis such that the program checks whether the substitution site is in spatial contact with critical residues annotated to be involved in forming binding sites or active sites. Additionally, the contacts of the substituted residue with ligands or subunits of the protein molecule are checked. After characterising the variant, PolyPhen applies empirically derived rules based on the characteristics to predict whether a missense variant is damaging or benign.

PolyPhen2

PolyPhen2 utilizes a combination of sequence- and structure-based attributes for the description of an amino acid substitution, and the effect of mutation is predicted by a naive Bayesian classifier. The sequence-based features include PSIC scores and MSA properties, and position of mutation in relation to domain boundaries as defined by Pfam [Finn et al., 2010]. The structure-derived features are solvent accessibility, changes in solvent accessibility for buried residues, and crystallographic B-factor.

SIFT

SIFT (Sorting Intolerant From Tolerant) makes inferences from sequence similarity using mathematical operations. SIFT constructs an MSA and considers the position of the missense variant and the type of the amino acid change. Based on the amino acids appearing at each position in the MSA, SIFT calculates the probability that a missense variant is tolerated conditional on the most frequent amino acid being tolerated.

SNAP

SNAP (Screening for Nonacceptable Polymorphisms) is a neural network-based tool for the prediction of the effect of a missense variant. The method utilises evolutionary information from PSI-BLAST [Altschul et al., 1997] frequency profiles and PSIC [Sunyaev et al., 1999], transition frequencies for mutations, biophysical characteristics of the substitution, secondary structural information, and relative solvent accessibility values predicted by PROFsec/PROFacc [Rost, 1996; Rost and Sander, 1994], chain flexibility predicted by PROFbval [Schlessinger et al., 2006], protein family evolutionary information, and information about domain boundaries from Pfam [Finn et al., 2010], and Swiss-Prot annotations [Bairoch and Apweiler, 2000] to classify a missense variant. The training sets for the NN were constructed from Protein Mutant Database (PMD) [Kawabata et al., 1999] data complemented by a set of neutral pseudomutations generated by the authors of the method as described in Bromberg and Rost [2007].

SNPs&GO

SNPs&GO is an SVM classifier based on mutation type and sequence environment information, sequence profiles taken from MSAs, predictions from the program Panther [Thomas et al., 2003], and a function-based log-odds score describing information about protein function defined by Gene Ontology (GO) terms [Ashburner et al., 2000].

From the output of the programs, we only took the binary prediction (pathogenic/neutral) into consideration without taking into account any confidence values provided by some of the programs. Panther provides a numerical output rather than a binary classification (subPSEC score), which we converted to a binary prediction using a cutoff point of -3 as recommended in [Thomas et al., 2003]. PolyPhen and PolyPhen2 classify the effects of a missense variant into three categories: “Probably pathogenic,” “Possibly pathogenic,” and “Benign.” We converted these into binary classifications in two ways, first by considering only the “Probably pathogenic” class as pathogenic and the “Possibly pathogenic” and “Benign” classes as neutral, and second, by considering both the “Probably pathogenic” and “Possibly pathogenic” classes as pathogenic, and the “Benign” class as neutral. These two ways of classifying the variants are referred to as PolyPhen(2)a and PolyPhen(2)b in this study, respectively.

Determination of Secondary Structural Elements and Accessible Surface Areas

The 3D structure coordinates of proteins were obtained from the PDB. Secondary structural information and accessible surface area (ASA) values for each mutation site were assigned by the program STRIDE [Frishman and Argos, 1995]. We classified residues with ASAs $\leq 10\%$ as buried and with ASAs $\geq 25\%$ as exposed, similarly as in a previous study [Khan and Vihinen, 2010].

Determination of Structural Classes of Proteins

The CATH database version 3.3 [Orengo et al., 1997] was used to group studied proteins according to their secondary and tertiary structure types.

Statistical Analyses

The quality of the predictions is described by six parameters: accuracy, precision, sensitivity, specificity, negative predictive value (NPV) and Matthews correlation coefficient (MCC). In the following equations, *tp, tn, fp*, and *fn* refer to the number of true positives, true negatives, false positives and false negatives, respectively.

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

$$\text{Precision} = \frac{tp}{tp+fp}$$

$$\text{Specificity} = \frac{tn}{fp+tn}$$

$$\text{Sensitivity} = \frac{tp}{tp+fn}$$

$$\text{NPV} = \frac{tn}{tn+fn}$$

$$\text{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp+fn)(tp+fp)(tn+fn)(tn+fp)}}$$

The MCC [Matthews, 1975] is a very important evaluation statistic as it is unaffected by the differing proportion of neutral and pathogenic datasets predicted by the different programs. Because of its insensitivity to differing test set sizes, it gives a more balanced assessment of performance than the other performance measures [Baldi et al., 2000].

To be able to correlate the quality parameters for different programs with different sizes of test sets containing different amounts of pathogenic and neutral cases, the numbers of neutral cases were normalized to be equal to the number of pathogenic cases for each program.

Substitution statistics for both the pathogenic and neutral datasets were analyzed by comparing the frequencies of the substitutions with the expected values that were calculated using the distribution of all amino acids in the datasets. For the original residues, the expected values were calculated with regard to their codon diversity thereby taking into account all possible amino acid substitutions. The chi-square test was used to determine the significance of the results and chi-square was calculated as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e is the expected frequency for an amino acid. The p-values were estimated in a one-tailed fashion.

Correlations between the program outputs were calculated by counting all of the common cases and those predicted correctly, and using Spearman’s rank correlation coefficient.

Results

Test Set Features

The distributions of mutated and mutant amino acids in both pathogenic and neutral datasets are biased (Table 2), and only a few residues occur as expected on the grounds of codon diversity. In the pathogenic dataset (mutation data), A, C, G, M, R, W, and Y are overrepresented among the original (mutated) amino acid residues, whereas E, F, I, K, L, N, Q, S, T, and V are significantly underrepresented. These results are in line with previous observations for distributions of disease-causing mutations in protein secondary structural elements [Khan and Vihinen, 2007], except for the overrepresentation of A and Y, and underrepresentation of L, N, S, and V in our data. In the neutral dataset, the distributions of many amino acids differ from the distributions in the pathogenic set. Most importantly, cysteines are highly underrepresented among the substituted positions, as opposed to their frequent mutation in the pathogenic dataset. This might be due to the important role of cysteines in folding of many proteins as they are capable of forming disulphide bonds, and therefore the substitution of cysteines in proteins transported through endoplasmic reticulum by any other residue can rarely be neutral in terms of protein structure and function. Other differences between the datasets are the underrepresentation of mutated glycine, tryptophan, and tyrosine residues in the neutral set as opposed to their frequent mutation in the pathogenic set, and the overrepresentation of isoleucine, asparagine, threonine, and valine residues in the neutral variation data, contrasting their underrepresentation in the mutation data.

The distributions of mutant or substituting amino acids are also very biased in both pathogenic and neutral datasets, and the amino acid residues I, P, R, T, V, and Y have opposite distributions in the mutation and neutral sets. Interestingly, proline residues are highly overrepresented among the substituting residues in the mutation dataset, and underrepresented in the negative set.

Table 2. Amino Acid Distributions in the Pathogenic (Mutations) and Neutral (SNPs) Datasets

	Wild-type residues/pathogenic variants					Wild-type residues/neutral variants			
	Observed	Expected	χ^2	*P*-value		Observed	Expected	χ^2	*P*-value
A	1224	252.5	**3737.28*****	0.000	A	1852	1449.4	**111.82*****	3.91E-26
C	943	468.1	**481.79*****	8.71E-107	C	424	473.9	*5.24**	0.022
D	950	988.7	1.52	0.218	D	991	1017.8	0.70	0.401
E	994	1449.8	*143.32****	5.02E-33	E	1273	1530.4	*43.31****	4.68E-11
F	537	766.1	*68.53****	1.25E-16	F	458	766.0	*123.83****	9.16E-29
G	2087	1355.0	**395.42*****	5.46E-88	G	1182	1374.1	*26.85****	2.20E-07
H	554	528.8	1.20	0.273	H	530	555.0	1.12	0.289
I	642	911.4	*79.64****	4.49E-19	I	996	924.5	**5.53***	0.019
K	497	1173.9	*390.28****	7.20E-87	K	774	1223.0	*164.85****	9.87E-38
L	1497	2068.4	*157.84****	3.35E-36	L	1270	2113.0	*336.34****	4.00E-75
M	520	435.5	**16.39*****	5.16E-05	M	642	442.2	**90.32*****	2.03E-21
N	605	754.4	29.59***	5.35E-08	N	894	777.0	**17.61*****	2.71E-05
P	1192	1252.8	2.95	0.086	P	1277	1323.3	1.62	0.203
Q	454	970.0	*274.52****	1.17E-61	Q	875	1028.1	*22.79****	1.81E-06
R	2797	1136.4	**2426.45*****	0.000	R	2376	1168.5	**1247.88*****	2.40E-273
S	1135	1681.4	*177.55****	1.66E-40	S	1648	1793.0	*11.72***	0.001
T	802	1087.9	*75.12****	4.42E-18	T	1482	1145.7	**98.72*****	2.91E-23
V	919	1246.3	*85.93****	1.86E-20	V	1682	1263.7	**138.46*****	5.78E-32
W	376	254.4	**58.17*****	2.41E-14	W	167	251.8	*28.54****	9.17E-08
Y	610	553.1	**5.85***	0.016	Y	377	549.8	*54.31****	1.71E-13
All	19335	19335			All	21170	21170		
	Mutant residues/pathogenic variants					Mutant residues/neutral variants			
A	622	1267.9	*329.01****	1.58E-73	A	1061	1388.20	*77.12****	1.61E-18
C	1233	563.5	**795.45*****	5.26E-175	C	722	617.0	**17.88*****	2.36E-05
D	900	633.9	**111.67*****	4.22E-26	D	666	694.1	1.14	0.286
E	719	563.5	**42.91*****	5.72E-11	E	825	617.0	**70.14*****	5.53E-17
F	623	633.9	0.19	0.664	F	855	694.1	**37.30*****	1.01E-09
G	922	1232.7	*78.29****	8.90E-19	G	1376	1349.6	0.52	0.473
H	918	633.9	**127.29*****	1.61E-29	H	967	694.1	**107.30*****	3.83E-25
I	619	950.9	*115.85****	5.14E-27	I	1139	1041.1	**9.20****	0.002
K	834	563.5	**129.85*****	4.41E-30	K	1171	617.0	**497.49*****	3.34E-110
L	1225	1796.1	*181.62****	2.15E-41	L	1390	1966.6	*169.06****	1.19E-38
M	534	317.0	**148.61*****	3.50E-34	M	828	347.0	**666.52*****	5.72E-147
N	662	633.9	1.24	0.265	N	845	694.1	**32.81*****	1.02E-08
P	1609	1267.9	**91.78*****	9.67E-22	P	1176	1388.2	*32.44****	1.23E-08
Q	808	563.5	**106.09*****	7.05E-25	Q	1056	617.0	**312.40*****	6.56E-70
R	2084	1831.4	**34.85*****	3.56E-09	R	1431	2005.2	*164.41****	1.23E-37
S	1502	1796.1	*48.17****	3.91E-12	S	1691	1966.6	*38.63****	5.13E-10
T	1012	1267.9	*51.64****	6.68E-13	T	1517	1388.2	**11.95****	0.001
V	1195	1267.9	*4.19**	0.041	V	1589	1388.2	**29.05*****	7.07E-08
W	638	246.5	**621.62*****	3.32E-137	W	471	269.9	**149.78*****	1.93E-34
Y	676	493.1	**67.88*****	1.74E-16	Y	394	539.9	*39.41****	3.44E-10
All	19335	19335			All	21170	21170		

The chi-square values in italics identify residues that are underrepresented and the values in bold identify overrepresented residues in comparison to random distributions derived theoretical codon usage frequencies. Significance levels are $^{*}P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$.

Proline is a known secondary structure breaker [Chou and Fasman, 1974] and therefore mutations to P are often pathogenic.

Performance of Prediction Methods

To evaluate the performance of the programs predicting the pathogenicity of missense variants, we used six measures: accuracy, precision (or positive predictive value, PPV), specificity, sensitivity, NPV, and MCC. The values for these measures are presented in Table 3 for all the missense variants. SNPs&GO performed best in terms of accuracy (0.82), precision (0.90), specificity (0.92), and MCC (0.65), but sensitivity was higher in six other methods, and MutPred, Panther, PolyPhen2b, and SNAP performed better in terms of NPV. nsSNPAnalyzer performed worst in terms of MCC (0.19), accuracy (0.60), NPV (0.60), and precision (0.59). The two versions of PolyPhen have very similar overall performance; however, PolyPhen2 is recommended because the quality measures are more balanced.. The version classifying "Probably pathonegenic," PolyPhen2a, as harmful is somewhat better than the other option.

In Table 3, the results are presented for the subset of cases for which structural information could be assigned. The performance of all methods was generally worse except for sensitivity, which is better for all methods. SNPs&GO performed best also in the structural subcategory considering accuracy, precision, specificity, and MCC, and MutPred was the best method in terms of sensitivity and NPV.

To test whether the poor performance was due to the smaller dataset size we sampled the full dataset results for those cases for which structural data was not available. We then compared the average MCC values of the samples to those obtained for the full dataset. The 100 sample datasets each contained randomly picked 1,000 neutral and 1,000 pathogenic variations. The average MCCs of the sample datasets were comparable to the MCCs of the full dataset in the case of Panther (average sample MCC 0.53), PhD-SNP (0.43), PolyPhen2b (0.39), and SNAP (0.47). For the other methods the MCC values were rather close when comparing the full dataset to the subdataset. We conclude that the large differences in the MCCs of the programs between the full dataset and the set for which structures were available (Table 3) were not due to the differences in the sizes of these datasets but were caused by some other factors, that is, differences in the performance of the methods when predicting on different types of data.

We also performed the analyses for a dataset that consisted only of LSDB-derived mutations not found in SwissProt (Table 3). This was done as some methods have been trained with Swiss-Prot disease-causing mutations. Because all methods (except SNPs&GO),

Table 3. Performance of Prediction Methods

	MutPred	nsSNPAnalyzer	Panther	PhD-SNP	PolyPhen1a	PolyPhen 1b	PolyPhen 2a	PolyPhen 2b	SIFT	SNAP	SNPs&GO
Performance of prediction methods (full data)											
tp[a]	13829	4360	9689	11900	10093	14285	13807	16206	10464	16000	13736
fn[a]	2507	2778	2859	6896	9185	4993	5102	2703	4856	2146	5487
tn[a]	15891	1319	8676	16788	17669	13671	13863	10199	12188	8190	17028
fp[a]	4557	943	2797	4377	3199	7197	6010	9674	7433	6387	1382
cases +[a]	16336	7138	12548	18796	19278	19278	18909	18909	15320	18146	19223
cases −[a]	20448	2262	11473	21165	20868	20868	19873	19873	19621	14577	18410
Accuracy[b]	0.81	0.60	0.76	0.71	0.69	0.70	0.71	0.69	0.65	0.72	0.82
Precision[b]	0.79	0.59	0.76	0.75	0.77	0.68	0.71	0.64	0.64	0.67	0.90
Specificity[b]	0.78	0.58	0.76	0.79	0.85	0.66	0.70	0.51	0.62	0.56	0.92
Sensitivity[b]	0.85	0.61	0.77	0.63	0.52	0.74	0.73	0.86	0.68	0.88	0.71
NPV[b]	0.84	0.60	0.77	0.68	0.64	0.72	0.72	0.78	0.66	0.83	0.76
MCC[b]	0.63	0.19	0.53	0.43	0.39	0.40	0.43	0.39	0.30	0.47	0.65
Performance of prediction methods (3D structure)											
tp[a]	5625	2857	3934	5041	4563	5980	5814	6726	4303	6751	5887
fn[a]	517	1603	1009	2411	3074	1657	1842	930	1329	714	1746
tn[a]	1101	569	735	1090	1361	1070	1163	843	904	700	1378
fp[a]	697	527	441	754	462	753	672	992	901	777	318
cases +[a]	6142	4460	4943	7452	7637	7637	7656	7656	5632	7465	7633
cases −[a]	1798	1096	1176	1844	1823	1823	1835	1835	1805	1477	1696
Accuracy[b]	0.76	0.58	0.71	0.63	0.67	0.68	0.70	0.67	0.63	0.69	0.79
Precision[b]	0.70	0.57	0.68	0.62	0.70	0.65	0.67	0.62	0.60	0.63	0.80
Specificity[b]	0.61	0.52	0.63	0.59	0.75	0.59	0.63	0.46	0.50	0.47	0.81
Sensitivity[b]	0.92	0.64	0.80	0.68	0.60	0.78	0.76	0.88	0.76	0.90	0.77
NPV[b]	0.88	0.59	0.75	0.65	0.65	0.73	0.72	0.79	0.68	0.83	0.78
MCC[b]	0.55	0.16	0.43	0.27	0.35	0.38	0.40	0.37	0.27	0.42	0.58
Performance of prediction methods (pathogenic dataset only from LSDBs, not in SwissProt)											
tp	2240	1175	1368	1436	1651	2410	2190	2764	2131	2615	2547
fn	899	862	1252	2158	1943	1184	1361	787	1145	917	952
tn	2655	212	1508	2842	3004	2333	2334	1705	2073	1382	2898
fp	804	165	501	752	534	1205	1028	1657	1268	1069	259
cases +[a]	3139	2037	2620	3594	3594	3594	3551	3551	3276	3532	3499
cases −[a]	3459	377	2009	3594	3538	3538	3362	3362	3341	2451	3157
Accuracy[b]	0.74	0.57	0.64	0.6	0.65	0.66	0.66	0.64	0.64	0.65	0.82
Precision[b]	0.75	0.57	0.68	0.66	0.75	0.66	0.67	0.61	0.63	0.63	0.90
Specificity[b]	0.77	0.56	0.75	0.79	0.85	0.66	0.69	0.51	0.62	0.56	0.92
Sensitivity[b]	0.71	0.58	0.52	0.4	0.46	0.67	0.62	0.78	0.65	0.74	0.73
NPV[b]	0.73	0.57	0.61	0.57	0.61	0.67	0.64	0.70	0.64	0.68	0.77
MCC[b]	0.48	0.14	0.28	0.21	0.33	0.33	0.31	0.30	0.27	0.31	0.66

[a] Total number of cases used by the given program (not normalized).
[b] Accuracy, precision, specificity, sensitivity, NPV, and MCC are calculated from normalised numbers.

and not only the ones trained on Swiss-Prot data, performed worse in this subcategory, we claim our results are not biased, even though we acknowledge that a perfectly fair comparison between methods trained on different datasets cannot be made.

To study the effect of residue types, the mutated and mutant amino acids were assigned into six groups according to their physicochemical properties: hydrophobic (C, F, I, L, M, V, W, and Y), positively charged (H, K, and R), negatively charged (D and E), conformational (G and P), polar (N, Q, and S), and A and T [Shen and Vihinen, 2004]. There were small differences in accuracy and precision of the methods for different types of wild-type or mutant amino acids, but their sensitivity and MCC were dependent on the physicochemical properties of the wild-type and mutant amino acids (Fig. 1). The methods were more sensitive to mutations at conformational, hydrophobic, and positively charged amino acids than mutations at polar residues or A and T (Fig. 1). MCC differed as well depending on the nature of the original residue position, and substitutions at hydrophobic positions were predicted best by most methods. Panther predicted mutations at hydrophobic and positively charged residues with equal performance, and MutPred and SNPs&GO performed better predicting conformational residues. Mutations affecting negatively charged residues had the lowest MCCs by most methods, except for PolyPhen1b, which predicted other classes better than the conformational class, and MutPred, nsSNPAnalyzer, and SNPs&GO, which had the lowest MCC when predicting the effects of mutations altering A and T residues (Fig. 1). The sensitivity and MCC of the methods also varied in predicting the effects of different types of mutant residues. All the methods performed best when the substituting residue was charged, and in the case of nsSNPAnalyzer, polar residues were predicted better than negatively charged residues, and SNAP predicted polar residues better than positively charged residues.

Differences in prediction sensitivity could also be seen at the level of individual amino acids. Predictions for substitutions at C, W, and Y were clearly more sensitive than at other residues by all methods (Fig. 2A). A similar trend was also seen when looking at mutant amino acids: mutations to the aforementioned residues were predicted with better sensitivity (Fig. 2A). The sensitivity of PolyPhen2b and SNAP varied less at individual residues than that of the other programs.

The results for the substitutions in the secondary structural elements are shown in Figure 2B. All of the programs predicted

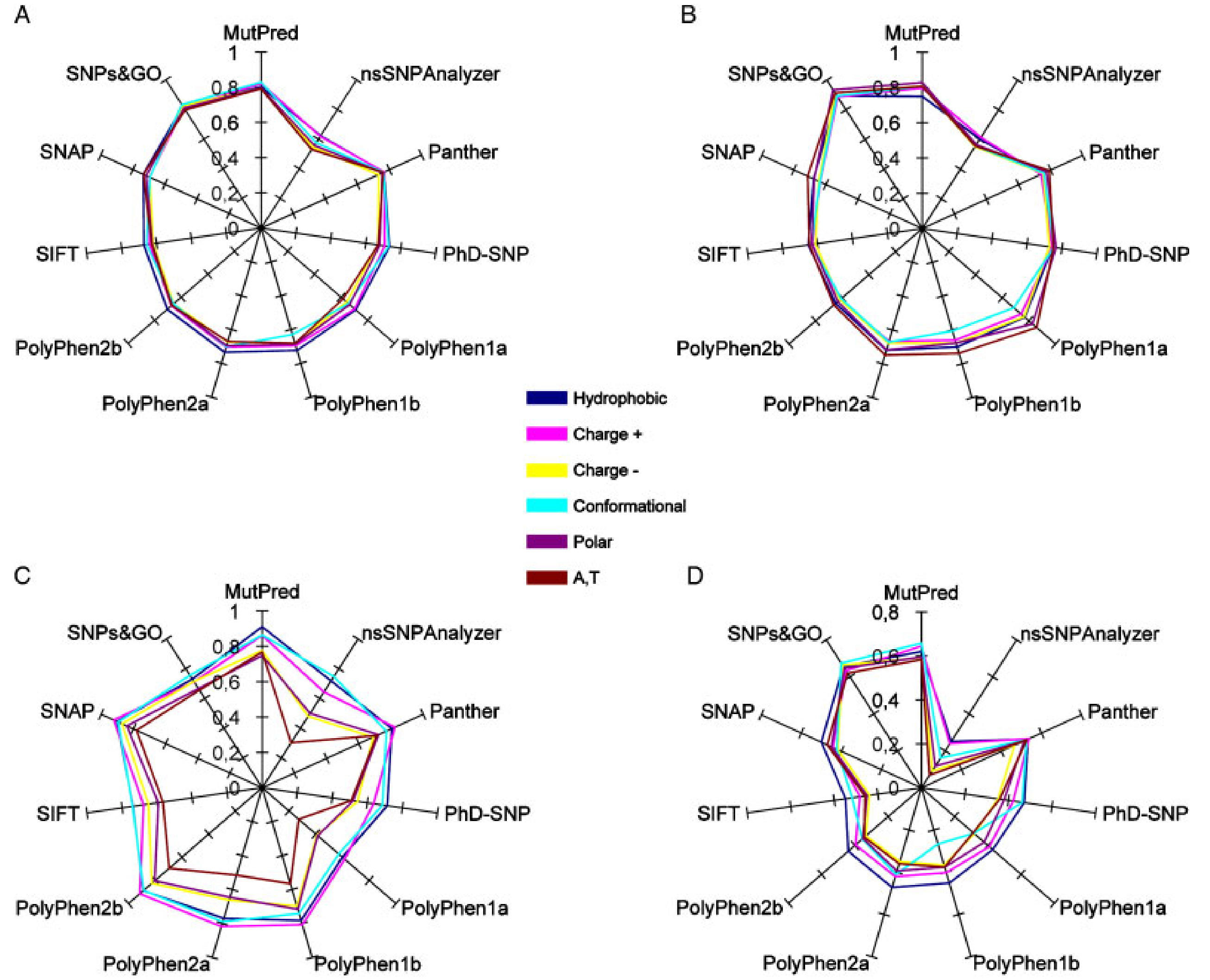

Figure 1. The values of the quality parameters, accuracy, precision, sensitivity, and Matthews correlation coefficient (MCC) for different classes of substituted amino acids. **A:** accuracy, **B:** precision, **C:** sensitivity, and **D:** MCC. Abbreviations: Charge+, positively charged. Charge −, negatively charged. [Color figures can be viewed in the online issue, which is available at wileyonlinelibrary.com]

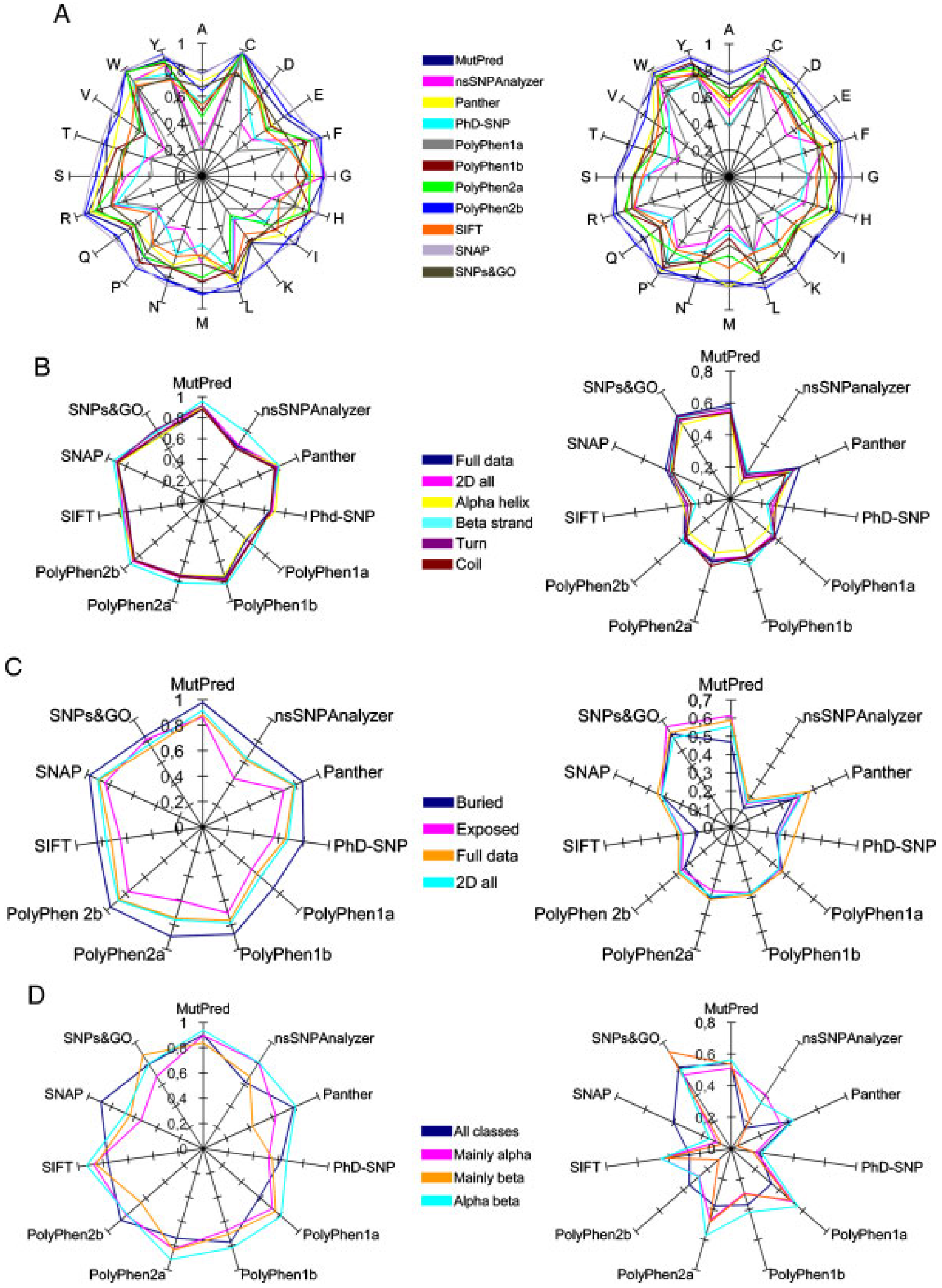

Figure 2. The values of sensitivity and Matthews correlation coefficient (MCC) for different types of amino acid substitutions. **A:** Sensitivity in different amino acid residues. Left: mutated (original) amino acids, right: substituting (mutant) amino acids. **B:** Sensitivity (left) and MCC (right) for amino acid substitutions at different secondary structural elements. **C:** Sensitivity (left) and MCC (right) for amino acid substitutions according to the accessible surface area (ASA) of the position (buried ASA ≤10%, exposed ASA ≥25%). **D:** Sensitivity (left) and MCC (right) for amino acid substitutions at different protein structural classes. [Color figures can be viewed in the online issue, which is available at wileyonlinelibrary.com]

the effects of substitutions at different secondary structures with almost equal accuracy and precision. Sensitivity and MCC values showed more variation with secondary structure. In terms of MCC, MutPred, nsSNPAnalyzer, PolyPhen1b, and PolyPhen2b predicted amino acid substitutions at strands best, whereas Panther, PolyPhen1a, SNAP, and SNPs&GO performed best at turns. PhD-SNP and SIFT predicted substitutions positioned at α-helices best, and PolyPhen2a at coils. The differences in MCC were not striking. Except for Panther, PhD-SNP, and SNPs&GO, all methods were most sensitive when predicting the effects of amino acid substitutions at strands. Solvent-accessible surface areas of the positions did not markedly affect prediction accuracy or precision, but all the methods were more sensitive when predicting the effects of substitutions at buried positions (Fig. 2C). MCC for most methods was better at exposed than buried positions, except for PolyPhen1a and PolyPhen2a, which performed better at buried positions. MCCs for PolyPhen1b and SNAP did not differ with solvent accessibility of the position. These results are not in line

Table 4. Pairwise Prediction Correlations

	MutPred	nsSNPAnalyzer	Panther	PhD-SNP	PolyPhen 1a	PolyPhen 1b	PolyPhen 2a	PolyPhen 2b	SIFT	SNAP	SNPs&GO
MutPred		8721	22645	36300	36522	36522	35198	35198	32705	29674	34066
nsSNPAnalyzer	4620		7237	9225	9380	9380	9353	9353	8270	8609	9145
Panther	15296	3589		23671	23869	23869	23406	23406	21540	20713	22555
PhD-SNP	23955	4389	14838		39659	39659	38254	38254	34532	32203	37095
PolyPhen1a	22125	4386	13961	22756		40146	38485	38485	34683	32533	37324
PolyPhen1b	22208	4965	14701	22170	23764		38485	38485	34683	32533	37324
PolyPhen2a	22234	4777	14728	21871	22383	23156		38782	33686	31790	36317
PolyPhen2b	20911	5012	14288	20042	19656	22412	24006		33686	31790	36317
SIFT	18807	4302	12623	18879	18207	18985	18645	17833		28726	32434
SNAP	18877	4750	13307	18004	17024	19811	19321	19945	16393		30987
SNPs&GO	23220	4672	14285	23333	22544	22206	22042	20569	18135	18833	
MutPred		53.0	67.5	66.0	60.6	60.8	63.2	59.4	57.5	63.6	68.2
nsSNPAnalyzer	0.36		49.6	47.6	46.8	52.9	51.1	53.6	52.0	55.2	51.1
Panther	0.54	0.37		62.7	58.5	61.6	62.9	61.0	58.6	64.2	63.3
PhD-SNP	0.57	0.35	0.51		57.4	55.9	57.2	52.4	54.7	55.9	62.9
PolyPhen1a	0.43	0.44	0.46	0.45		59.2	58.2	51.1	52.5	52.3	60.4
PolyPhen1b	0.43	0.47	0.50	0.43	0.66		58.2	58.2	54.7	60.9	59.5
PolyPhen2a	0.49	0.44	0.51	0.45	0.56	0.58		61.9	55.3	55.3	60.7
PolyPhen2b	0.44	0.42	0.49	0.40	0.46	0.57	0.72		52.9	62.7	56.6
SIFT	0.41	0.53	0.48	0.45	0.45	0.52	0.50	0.51		57.0	55.9
SNAP	0.46	0.41	0.51	0.44	0.44	0.54	0.52	0.53	0.53		60.8
SNPs&GO	0.50	0.25	0.39	0.44	0.39	0.38	0.38	0.34	0.34	0.39	

Upper table: the number of cases shared by two programs (upper right triangle). The number of cases predicted correctly (lower left triangle). Lower table: The number of cases predicted correctly, reported as a percentage (upper right triangle). Pairwise correlation (lower left triangle).

with a previous study [Mort et al., 2010], where a sequence conservation based method yielded results of lower accuracy when predicting the effects of solvent-exposed residues.

CATH classifies proteins as mainly α-helical or β-stranded, mixed α- and β-structures (α–β), or as having few secondary structures. Interestingly, none of the proteins included in this analysis was assigned into the few secondary structures class. The predictions differed with respect to sensitivity and MCC depending on which protein class a mutation appeared (Fig. 2D). Most programs were more sensitive to amino acid substitutions in the α–β class of proteins, but SNPs&GO predicted substitutions best in the mainly β-class. nsSNPAnalyzer predicted those mutations occurring in α–β and α-helical proteins or domains with equal sensitivity. MCCs varied significantly with the structural class of proteins, especially in the predictions by nsSNPAnalyzer, PolyPhen1b, PolyPhen2a, and 2b, and SNPs&GO. The results were generally better for the α–β class of proteins, but nsSNPAnalyzer predicted substitutions at α-helical proteins best and SNPs&GO performed best with proteins in the mainly β-class.

To further evaluate the performance of the programs we compared them in a pairwise fashion (Table 4). The numbers of cases that were shared by the programs varied because the number of cases that could be predicted by each program varied as described in the Materials and Methods section. The largest percentage of correctly predicted cases by two programs was 68.2% (for the combination of MutPred and SNPs&GO). On average, the fraction of correctly predicted cases between any two programs was 57.7%. The correlations between two programs were highest for MutPred and PhD-SNP (0.57), and for PolyPhen 1 and 2 (0.57 for the less stringent b versions, and 0.56 for the a versions) (without taking into account the higher correlation between PolyPhen1a or 2a and PolyPhen1b or 2b that are different forms of the same program). Correlation was lowest for nsSNPAnalyzer and SNPs&GO (0.25).

Discussion

In this study we evaluated how reliably the pathogenicity of missense mutants can be predicted, and whether selected features of the variant or the structural context affect prediction performance. The processing of the vast and increasing amount of genetic variation data requires the development of automatic annotation tools to determine the potential pathological character of a given variant. Prioritizing the most interesting and likely pathogenic cases for experimental analysis is another important application of the tested prediction methods.

To our knowledge, no comprehensive evaluation of the performance of missense variant pathogenicity predictors has been made outside the performance studies of individual methods in the context of their development. We selected test sets that have not been used in the training of the methods as such, but a subset of the pathogenic dataset is comprised of mutations from Swiss-Prot, and some methods (MutPred, nsSNPAnalyzer, PhD-SNP, PolyPhen2, and SNPs&GO) have used Swiss-Prot mutations in the training of the method. Testing of the performance of a method with the same cases it was trained on would lead into biased results, so that those methods trained on SwissProt mutations would have an advantage over the other methods. However, because the pathogenic dataset includes a large number of LSDB variations not found in SwissProt, we claim the test set was not similar to the training sets to the extent that it would advantage those methods trained on SwissProt data. Further, we tested the methods with cases coming only from LSDBs. With this dataset the performance decreased with all methods, whether trained on Swiss-Prot data or not, except for SNPs&GO. This indicates that the good performance of SNP&GO was not a result of that it has previously been exposed to the test dataset during its training phase. Furthermore, the poor performance of PhD-SNP indicates the method did not benefit from the possible identical cases in the data used for training and testing. However, it is impossible to construct a large testing dataset that would not share any cases with the original training sets of any of the methods, especially when the specific contents of the training sets are rarely published.

The neutral dataset was generated from dbSNP entries that had a frequency higher than 1% when there was data at least for 25 individuals (50 chromosomes). This way the number of false negatives could be minimized in the test set.

There are still other pathogenicity predictors that we did not evaluate. SNPs3D [Yue et al., 2006] was not included in this study because it does not allow submission of user-defined amino acid substitutions. Similarly, LS-SNP [Karchin et al., 2005] is an annotated database of SNPs, not a prediction method for any user-provided variant, although often referred to as a prediction method for nsSNP pathogenicity. The Auto-Mute predictor of disease potential of human nsSNPs [Barenboim et al., 2008] was left out from the analysis because the program did not allow batch submission. PMut [Ferrer-Costa et al., 2005] could not be tested because the server did not return predictions.

Overall, we found SNPs&GO and MutPred to be clearly the most reliable predictors for our dataset of genetic variants. The accuracies of all the methods were in the range of 0.60–0.82, and precision ranged from 0.59 to 0.90. More variation among the methods was seen when considering the sensitivities and MCC values that ranged from 0.52 to 0.88 and 0.19 to 0.65, respectively. The local structural context of a mutated residue did not dramatically affect predictor performance in most cases but most methods showed variance in their prediction power at the level of protein tertiary structure classification and at different mutated positions.

Studies have shown that combining information obtained from the multiple sequence alignment and three-dimensional protein structure can increase prediction performance [Bromberg and Rost, 2007; Saunders and Baker, 2002]. According to our results, this is not always the case. Panther operates solely on sequence-based evolutionary information, and it is one of the best performing methods, outperforming all the methods incorporating structural information in the prediction, except for MutPred, which uses sequence-derived structural predictions as features in combination with evolutionary information. Furthermore, although nsSNPAnalyzer uses the SIFT method for the evolutionary analysis and also includes structure-derived features, its overall performance is below that for SIFT, except for an increase in specificity in the structure subset of data. However, the two best performing predictors include both protein structural or functional and MSA-derived information in the prediction.

It is very difficult to determine whether the notable differences in the performance of these methods are caused by differences in the features utilized by the methods or the training datasets. For example, SNPs&GO uses GO annotations as a feature, and GO is biased toward genes involved in diseases. The PDB is biased as well, containing structures of mostly well-studied proteins, which include products of disease-related genes. Therefore, one would expect SNPs&GO would perform better in predicting the effects of missense variants in proteins that have structures in the PDB as they are likely to have GO annotation as well—and in fact, it performs worse. One factor that very probably affects prediction reliability is the quality of multiple sequence alignment. Because all of the methods studied here use MSA as input to the prediction, the quality of the provided MSA should be very carefully assessed. For many of the methods, we did not find documentation how the MSA is constructed when the user provides just the query sequence as input. For example, an automatic BLAST search often performed by the programs may lead into construction of an MSA that contains multiple versions of the same sequence or paralog sequences, affecting the resulting conservation analysis. The MSA should contain a selection of closely and distantly related sequences in order to effectively yield a conservation signal.

In conclusion, those methods that performed best had high accuracy (reaching 0.82, SNPs&GO), precision (0.90, SNPs&GO), specificity (0.92, SNPs&GO), sensitivity (0.88, SNAP), and NPV (0.84, MutPred). Matthews correlation coefficient reached the value of 0.65 at best (SNPs&GO). There is no single method that could be rated as best by all parameters, so the user should consider what aspects would be most valuable considering the nature of the data analysed. Furthermore, some methods require 3D structure coordinates, limiting the number of cases that can be analyzed (nsSNPAnalyzer), and some methods are at least currently too slow for high-throughput analyses (SNAP). Although some of the existing methods perform reasonably well, development of new, more reliable methods is certainly needed. Complementary methods could be combined in a metaserver to yield more reliable predictions.

Acknowledgments

The authors thank Pier Luigi Martelli and Rita Casadio from the University of Bologna, Biao Li from Indiana University, and Sean Mooney and Vidhya Krishnan from the Buck Institute for Age Research, for cooperation in running of data.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–449.

Ahmad S, Gromiha MM, Sarai A. 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48.

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated non-synonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482.

Barenboim M, Masso M, Vaisman, II, Jamison DC. 2008. Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. Proteins 71:1930–1939.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306–W310.

Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. 2004. The ASTRAL Compendium in 2004. Nucleic Acids Res 32:D189–D192.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry 13: 222–245.

Daily KM, Radivojac P., Dunker AK. 2005. Intrinsic disorder and protein modifications: building an SVM predictor for methylation. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB: 475–481.

Delorenzi M, Speed T. 2002. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18:617–625.

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21:3176–3178.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res 3:D211–D222.

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. Proteins 23:566–579.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, and others. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28:554–562.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32:1037–1049.

Johnson GC, Todd JA. 2000. Strategies in complex disease mapping. Curr Opin Genet Dev 10:330–334.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.

Karchin R. 2009. Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Kawabata T, Ota M, Nishikawa K. 1999. The protein mutant database. Nucleic Acids Res 27:355–357.

Khan S, Vihinen M. 2007. Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7:56.

Khan S, Vihinen M. 2010. Performance of protein stability predictors. Hum Mutat 31:675–684.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. Science 252:1162–1164.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451.

McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. J Mol Biol 238:777–793.

Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6:44–56.

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters B, Sathyesh R, Li B, Sun Y, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P, Mooney SD. 2010. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335–346.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80.

Ng PC, Henikoff JG, Henikoff S. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16:760–766.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. Structure 5: 1093–1108.

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

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27:1200–1208.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. 2004. Protein flexibility and intrinsic disorder. Protein Sci 13: 71–80.

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. 2010. Identification, analysis, and prediction of protein ubiquitination sites. Proteins 78:365–380.

Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. 2006. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins 63:398–410.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900.

Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent WJ. 2011. 0ENCODE whole-genome data in the UCSC Genome Browser (2011 update). Nucleic Acids Res 39(Database issue):871–875.

Rost B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525–539.

Rost B, Sander C. 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891–901.

Schlessinger A, Yachdav G, Rost B. 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22:891–893.

Shen B, Vihinen M. 2004. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. Protein Eng Des Sel 17:267–276.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505–513.

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12: 387–394.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30: 703–714.

Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29:361–366.

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. 2004. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23: 464–470.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.