Bioinformatics, 31(10), 2015, 1599–1606
doi: 10.1093/bioinformatics/btu862
Advance Access Publication Date: 7 January 2015
Original Paper

OXFORD

Genetics and population analysis

DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels

Lukas Folkman[1,2,3], Yuedong Yang[1,2,4], Zhixiu Li[1,2,4], Bela Stantic[1,2], Abdul Sattar[1,2,3], Matthew Mort[5], David N. Cooper[5], Yunlong Liu[6] and Yaoqi Zhou[1,2,4,*]

[1]School of Information and Communication Technology, Griffith University, Parklands Drive, Southport, Queensland 4222, Australia, [2]Institute for Integrated and Intelligent Systems, Griffith University, 170 Kessels Road, Brisbane, Queensland 4111, Australia, [3]Queensland Research Laboratory, NICTA – National ICT Australia, 70-72 Bowen Street, Spring Hill, Queensland 4000, Australia, [4]Institute for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 4222, Australia, [5]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK and [6]Department of Medical and Molecular Genetics, Indiana University School of Medicine, 975 West Walnut Street, MRL Bldg IB130, Indianapolis, IN 46202, USA

*To whom correspondence should be addressed.

Associate Editor: Gunnar Ratsch

Received on October 3, 2014; revised on December 4, 2014; accepted on December 23, 2014

Abstract

Motivation: Frameshifting (FS) indels and nonsense (NS) variants disrupt the protein-coding sequence downstream of the mutation site by changing the reading frame or introducing a premature termination codon, respectively. Despite such drastic changes to the protein sequence, FS indels and NS variants have been discovered in healthy individuals. How to discriminate disease-causing from neutral FS indels and NS variants is an understudied problem.

Results: We have built a machine learning method called DDIG-in (FS) based on real human genetic variations from the Human Gene Mutation Database (inherited disease-causing) and the 1000 Genomes Project (GP) (putatively neutral). The method incorporates both sequence and predicted structural features and yields a robust performance by 10-fold cross-validation and independent tests on both FS indels and NS variants. We showed that human-derived NS variants and FS indels derived from animal orthologs can be effectively employed for independent testing of our method trained on human-derived FS indels. DDIG-in (FS) achieves a Matthews correlation coefficient (MCC) of 0.59, a sensitivity of 86%, and a specificity of 72% for FS indels. Application of DDIG-in (FS) to NS variants yields essentially the same performance (MCC of 0.43) as a method that was specifically trained for NS variants. DDIG-in (FS) was shown to make a significant improvement over existing techniques.

Availability and implementation: The DDIG-in web-server for predicting NS variants, FS indels, and non-frameshifting (NFS) indels is available at http://sparks-lab.org/ddig.

Contact: yaoqi.zhou@griffith.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A key requirement for personalized medicine is to fully annotate human genetic variations in different individuals. Some genetic variations are benign, while others are disease-causing or disease-associated. In addition to single nucleotide variations that change a single amino acid in a protein sequence (missense mutations), a large fraction of pathological human genetic variations are caused by nonsense (NS) mutations [comprising 11% (Stenson *et al.*, 2014)] where a single nucleotide variation introduces a premature termination codon (PTC), and by microinsertions and microdeletions [*indels*, comprising 24% (Ball *et al.*, 2005)] that involve the insertion or deletion of ≤20 nucleotides. In exonic (protein-coding) regions, indels can be frameshifting (FS) or *non*-frameshifting (NFS) depending upon whether or not the indel in question inserts or deletes a multiple of three nucleotides. NFS indels insert/delete multiples of three nucleotides and hence do not alter the coding region, comprising three-nucleotide codons, other than at the indel site. On the other hand, FS indels, having a length indivisible by three, shift the reading frame and alter the coding sequence downstream of the indel site, which depending on the location and downstream sequence context of the indel may introduce a PTC. Thus, both FS indels and NS variants alter the entire coding sequence downstream from the variation site. As a result, they are often assumed to have a significant functional impact and to be potentially disease-causing. Indeed, NS variants account for ∼20% of all disease-associated single base substitutions (Mort *et al.*, 2008). However, a considerable number of NS variants and FS indels have been identified as being benign in recent studies (McVean *et al.*, 2010; Mills *et al.*, 2011). How to distinguish neutral from potentially disease-causing FS indels and NS variants is therefore of both practical and fundamental interest.

The functional effects of FS indels and NS variants are poorly understood although variations which introduce a PTC and activate nonsense-mediated decay (NMD) are more likely to come to clinical attention (Mort *et al.*, 2008). Additionally, FS indels and NS variants may disrupt pre-mRNA splicing with around ∼31% of disease-causing NS variants predicted to disrupt splicing (Mort *et al.*, 2014). It is therefore clear that the functional impact of this class of variants (FS indels and NS mutations) is not always straightforward to interpret and currently most available methods for discriminating deleterious and neutral genetic variants are devoted exclusively to missense mutations. Examples are SIFT (Ng and Henikoff, 2001), PolyPhen (Adzhubei *et al.*, 2010) or MutPred (Li *et al.*, 2009), for recent reviews see Thusberg *et al.* (2011) or Bendl *et al.* (2014). Although there are several methods for NFS indels [DDIG-in (NFS) (Zhao *et al.*, 2013), SIFT Indel (Hu and Ng, 2013), PROVEAN (Choi *et al.*, 2012), PinPor (Zhang *et al.*, 2014) and KD4i (Bermejo-Das-Neves *et al.*, 2014)], there are only three methods available for FS indels and two for NS variants. D-score (Zia and Moses, 2011) ranks FS indels and NS variants based on the loss of protein information content derived from the conservation of the target protein sequence without specific training. SIFT Indel (Hu and Ng, 2012) is a decision tree trained on disease-causing FS indels annotated in the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2014) and neutral indels from pairwise alignments between human proteins and their functional orthologs in cow, dog, horse, chimpanzee, rhesus macaque and rat. SIFT Indel employs four features: fraction of affected conserved DNA bases, relative indel location, fraction of affected conserved amino acids, and indel distance to the exon boundary. CADD (Combined Annotation-Dependent Depletion) (Kircher *et al.*, 2014) is a general framework for predicting all possible types of genetic variations. It is based on a support vector machine (SVM) model trained with a variety of features including scores calculated with other methods such as SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei *et al.*, 2010). The training neutral dataset was compiled from variants between human and inferred human-chimpanzee common ancestral genomes. The training deleterious dataset was created using a genome-wide simulator of *de novo* germline mutations.

In this article, we have built the first machine learning technique for FS indels and NS variants trained entirely on actual human variants, rather than on simulated deleterious variants (CADD) or variants derived from orthologous animal proteins (SIFT Indel). Moreover, we investigated the application of structural properties, which have been shown to be important for the classification of NFS indels (Zhao *et al.*, 2013) and missense mutations (Adzhubei *et al.*, 2010; Li *et al.*, 2009). By utilizing disease-causing variants annotated in the HGMD (Stenson *et al.*, 2014) and putatively neutral variants from the 1000 GP (McVean *et al.*, 2010), we found that the most discriminative feature for FS indels and NS variants was the disruption of DNA conservation, rather than the disruption of protein structure as in the case of NFS indels. The best combination of eight features identified by a feature selection algorithm was used to build the final SVM model named DDIG-in (FS) (Detecting DIsease-causing Genetic variations). More importantly, the method developed here was subjected to rigorous independent testing. These tests included the use of FS indels derived from functional animal orthologs as a neutral dataset (the SIFT Indel neutral dataset) and the use of NS variants for testing the method trained on FS indels. Here, we have rationalized the possible equivalence between FS indels and NS variants by their effect on the protein sequence: FS indels effectively render the protein sequence meaningless after the indel site, and probably have the same effect as truncation of the protein induced by a NS variant. DDIG-in (FS) achieved a Matthews correlation coefficient (MCC) of 0.59 and 0.54 for the 10-fold cross-validation and independent test (on non-overlapping HGMD disease-causing indels and neutral indels from the SIFT Indel dataset), respectively, compared with an MCC of 0.38 and 0.35 for CADD. A very different performance, MCC of 0.29 and 0.63 for the two different datasets, was observed for SIFT Indel. The robust performance of DDIG-in (FS) was further confirmed by an independent dataset of NS variants and negative correlation with average allele frequency (AF) in the presumably healthy population from the 1000 GP.

2 Methods

2.1 Datasets

We describe the datasets used in this work below. For details on how we compiled these datasets, refer to Supplementary Methods. The dataset of inherited disease-causing and putatively neutral FS indels in coding regions was retrieved from the HGMD (version Professional 2012.2) (Stenson *et al.*, 2014) and 1000 GP (phase 1, version 3, 20101123) (McVean *et al.*, 2010), respectively. The final *FS indels dataset* comprised 660 disease-causing and 580 neutral

indels in 660 and 491 protein-coding genes, respectively. We also built an independent test set from non-overlapping disease-causing HGMD indels and neutral indels from the SIFT Indel training dataset, which was derived from functional animal orthologs (Hu and Ng, 2012). The final *HGMD + SIFT FS indels dataset* contained 2008 disease-causing and 2008 neutral FS indels in 737 and 1996 genes, respectively. Importantly, the HGMD + SIFT FS indels dataset had a sequence similarity ≤30% to the FS indels dataset. Another independent test set comprised NS variants derived from the HGMD (disease-causing) and 1000 GP (neutral). The final *NS variants dataset* contained 3861 disease-causing and 3861 neutral variants in 1122 and 2989 genes, respectively. Also the NS variants dataset had a sequence similarity ≤30% to the FS indels dataset. Table 1 depicts how these three datasets were utilized for training and testing of our two distinct methods [DDIG-in (FS) and DDIG-in (NS)].

2.2 Machine learning features

We evaluated 36 different features (see Supplementary Methods and Supplementary Table S1) that could be used for discriminating between disease-causing and neutral FS indels as well as NS variants. We used the sequential forward floating selection (SFFS) algorithm (Pudil *et al.*, 1994) to find an effective combination of features which we used for training the final SVM model (Cortes and Vapnik, 1995). Details on how we performed the feature selection and optimized the SVM model can be found in Supplementary Methods.

2.2.1 Global features

We considered four global *gene-level* features. Feature K_a/K_s *ratio* was calculated as the ratio of the number of non-synonymous (K_a) and synonymous (K_s) substitutions in a given gene. K_a/K_s ratio is commonly used as an indicator of selective pressure acting on a protein-coding gene (Hurst, 2002). We downloaded pre-calculated K_a and K_s values for the human-chimpanzee, human-macaque, human-mouse and human-rat alignments from Ensembl (release 75) (Flicek *et al.*, 2013) and calculated the average K_a/K_s ratio for the four alignments.

As a gene may contain exons separated by introns, we examined the following features for evidence of a variant affecting alternative splicing: *number of transcripts*, *fraction of unaffected transcripts* and *fraction of translatable transcripts*. Here, an unaffected transcript (splice isoform) is one for which the variation is located within an intron (while the variation is located in an exon for some other splice isoform of the same gene). The number of translatable transcripts is given as the number of transcripts which would *not* be candidates for non-stop decay (NSD) or NMD induced by the genetic variation. NSD and NMD are mRNA surveillance mechanisms which prevent translation of damaged genes. We considered a variant to be a candidate for NSD if it lacked a PTC (van Hoof *et al.*, 2002). A variant was considered to be a NMD candidate if the genetic variation occurred more than 49 nucleotides (discounting intervening introns) from the 3′-most intron (Nagy and Maquat, 1998).

We also examined nine global *transcript-level* features (calculated as the average, minimum or maximum of all transcripts of the given gene): *number of exons*; *relative exon number*; *distance to the nearest upstream* (*downstream*) *splice site*; *relative distance to the 5′ end*, *3′ end* and *centre of the sequence*; *relative length of the variant*; and *relative length of the mutated sequence*. The details on how we calculated these features can be found in Supplementary Methods.

Table 1. Three datasets employed for training and testing DDIG-in (FS indels) and DDIG-in (NS variants)

Dataset	Disease	Neutral	DDIG-in	
	(HGMD)	(GP/SIFT)	(FS)	(NS)
FS indels	660	580 (GP)	training, CV	—
HGMD + SIFT FS indels	2008	2008 (SIFT)	testing	—
NS variants	3861	3861 (GP)	testing	training, CV

We ensured that the two datasets used for testing DDIG-in (FS) were mutually independent from the DDIG-in (FS) training dataset by removing genes with protein sequences that had a pairwise sequence identity >30%.

HGMD: Human Gene Mutation Database; GP: 1000 GP; SIFT: neutral indels derived from animal orthologs; CV: 10-fold cross-validation.

2.2.2 Local features

With the exception of a *nucleotide-level* feature *DNA conservation*, all other local features were *protein-level* features. The DNA conservation was derived from phylogenetic *P*-values of the multiple alignments of 45 vertebrates to the human genome calculated with the phyloP program (Pollard *et al.*, 2010). We downloaded pre-calculated phyloP scores from the UCSC Table Browser (Karolchik *et al.*, 2004).

On the *protein-level*, we used PSI-BLAST (Altschul *et al.*, 1997) (NCBI non-redundant database, three iterations, *e*-value threshold of 0.001) to create a position-specific scoring matrix (PSSM) and weighted observed percentages matrix from which we calculated feature *PSSM conservation* (the equation is given in Supplementary Methods).

We also considered six entropy-based protein conservation scores which were studied and implemented by Capra and Singh (2007). The best performance using cross-validation on the training set was achieved by Shannon entropy (SE) calculated from the 30 most related sequences (ranked by the *e*-value threshold) from the multiple sequence alignment generated with PSI-BLAST (see above). The feature *protein conservation* (*SE*) was scaled to fit in the range (0, 1) where 1 is the highest protein conservation.

Ten other protein conservation features were calculated using HHblits (Remmert *et al.*, 2012) from hidden Markov model sequence profiles. Details about these features are provided in Supplementary Methods.

Finally, we considered five structural features predicted from the protein sequence: *relative accessible surface area*, *helix*, *sheet*, *coil* and *disorder probabilities*. The accessible surface area and secondary structure probabilities were all predicted using SPINE-X (Faraggi *et al.*, 2009, 2011). The disorder probability was calculated using SPINE-D (Zhang *et al.*, 2012).

2.2.3 Feature extraction of local features

Whereas the global features were calculated as a single-valued property of the whole coding sequence or the protein product, local features quantified a given property over a *window* of neighbouring coding bases (residues) of the given gene (protein). The window encompassed all bases (residues) from the variation site to the 3′ end (C-terminus) plus *n* bases (residues) from the variation site towards the 5′ end (N-terminus). We refer to this window as the *lost-sequence* window. We also considered two other types of windows, *small-symmetric* and *next-splice* (see Supplementary Methods), but they yielded a lower performance as evaluated with one nucleotide-level feature (DNA conservation) and one protein-level feature

(PSSM conservation). We used the same two features to optimize the window length for the lost-sequence window. The best performance using cross-validation on the training set was achieved with $n=40$ bases for DNA conservation and $n=5$ residues for PSSM conservation. We kept the window length fixed for all other features in all experiments.

Each local feature was encoded as average, minimum, or maximum within the lost-sequence window. In addition, we also calculated the number of bases (residues) for which the value of the feature was higher than a threshold t within the given window relative to the number of bases (residues) in the given window (H^t) or relative to the number of bases (residues) with feature values $>t$ for the entire nucleotide (protein) sequence (H_g^t). Thus, a total of five *encoding methods* denoted as avg, min, max, H^t and H_g^t. Threshold t was optimized for each local feature individually (Supplementary Table S1).

2.3 Evaluation

We used 10-fold cross-validation on the FS indels dataset to design our method, select relevant features, and optimize all parameters (window type, window length, SVM parameters, etc.). To avoid over-training on specific genes, we ensured that no two cross-validation folds shared similar sequences. This means that all indels of any cluster of similar sequences were contained within a single fold. The clusters were determined with protein-level Blastclust (Altschul *et al.*, 1990) with a sequence similarity threshold of 30%. In addition, we kept the ratio of disease-causing and neutral indels reasonably similar between the folds. We replicated our experiments 100 times with randomly re-generated folds, averaged the results and calculated standard deviations. Because we had an abundance of disease-causing variants from the HGMD (see Supplementary Methods), each of the 100 replications used a different random sample of disease-causing FS indels. We employed the HGMD + SIFT FS indels and the NS variants datasets for independent testing of our method. Importantly, these datasets did not share sequences with >30% sequence similarity with the training FS indels dataset. Again, we replicated all experiments 100 times with random sampling of the disease-causing variants.

We assessed the overall prediction performance in terms of the receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC). An ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1−specificity) at different prediction thresholds. Furthermore, we calculated MCC, binary classification accuracy (Q_2), sensitivity (Se, recall), specificity (Sp), positive predictive value (PPV, precision), and negative predictive value (NPV). The equations of these evaluation measures are given in Supplementary Methods.

3 Results

3.1 Single feature performance on FS indels dataset

We first examined the ability of each of the 36 individual machine learning features (Supplementary Table S1) to discriminate between the disease-causing and neutral FS indels. Table 2 ranks the 10 best performing features and their best performing encoding methods according to the AUC along with other measures including MCC and Q_2. The top three most discriminative features were the fraction of highly conserved DNA positions with an AUC of 0.83, followed by the minimum protein conservation calculated as Shannon entropy (SE) (AUC of 0.74), and the ratio of the number of non-synonymous (K_a) and synonymous (K_s) substitutions in a given gene (AUC of 0.74). Other important features were sequence conservation scores based on the PSSM and hidden Markov model profiles from HHblits, features related to RNA splicing (number of exons and relative exon number), protein structural features (predicted coil probability and relative accessible surface area), and the fraction of translatable transcripts.

Figure 1 depicts the distributions of the top three features [DNA conservation, protein conservation (SE), and K_a/K_s ratio] for disease-causing and neutral FS indels. Disease-causing indels occurred more frequently within regions with higher fractions of conserved DNA bases and with higher minimal protein conservation scores and in genes with low K_a/K_s ratios (i.e. fewer mutations that change amino acid residues). The results summarized in Table 2 and Figure 1 demonstrate that sequence conservation in different forms and at different molecular levels (DNA or protein) yielded the best discrimination of disease-causing FS indels.

3.2 Feature selection and 10-fold cross-validation with FS indels dataset

We employed SVM and a greedy SFFS algorithm to select a well-performing subset from 54 diverse feature + encoding pairs (Supplementary Table S2). SFFS selected eight features (Supplementary Table S3) including the three best features [DNA conservation, protein conservation (SE), and K_a/K_s ratio] from our single-feature experiment (previous section). Four additional features were related to the top 10 features from Table 2 [PSSM conservation (average and fraction of highly conserved residues), and fractions of translatable and unaffected transcripts]. One new feature was the minimum of the predicted disorder probability within the lost-sequence window. Most of these features reflected either the local conservation in sequence (DNA and protein) and structure (protein), or the global conservation of a gene (K_a/K_s ratio). For convenience, we refer to the final SVM model encompassing these eight features as the DDIG-in (FS) method.

DDIG-in (FS) yielded an AUC of 0.87 and MCC of 0.59 (Table 3). Sensitivity and specificity reached 86 and 72%, respectively. Whereas DNA conservation performed quite well on its own, DDIG-in (FS) still yielded an absolute (relative) improvement of 0.04 (5%) for AUC and 0.05 (9%) for MCC. Figure 2a compares the performance of DDIG-in (FS) with the performance of the top three features in terms of ROC curves. We also compared the performance of our method with two other available methods: SIFT Indel (Hu and Ng, 2012) for predicting damaging FS indels and

Table 2. Top 10 discriminative features for the FS indels dataset

Feature (encoding method[a])	Type	AUC[b]	MCC[b]	Q_2[b]
DNA conservation ($H^{2.5}$)	local	0.83	0.54	0.77
protein conservation (SE[c]) (min)	local	0.74	0.39	0.69
K_a/K_s ratio	global	0.74	0.40	0.70
number of exons (min)	global	0.68	0.32	0.63
PSSM conservation (min)	local	0.67	0.30	0.65
HHblits match to deletion (avg)	local	0.67	0.29	0.65
coil probability (max)	local	0.65	0.23	0.60
relative exon number (min)	global	0.65	0.36	0.68
relative accessible surface area (min)	local	0.65	0.24	0.62
fraction of translatable transcripts	global	0.64	0.29	0.65

[a]For definitions of the encoding methods see Section 2.

[b]AUC, MCC and Q_2 are the area under the ROC curve, Matthews correlation coefficient and binary classification accuracy, respectively.

[c]SE denotes Shannon entropy.

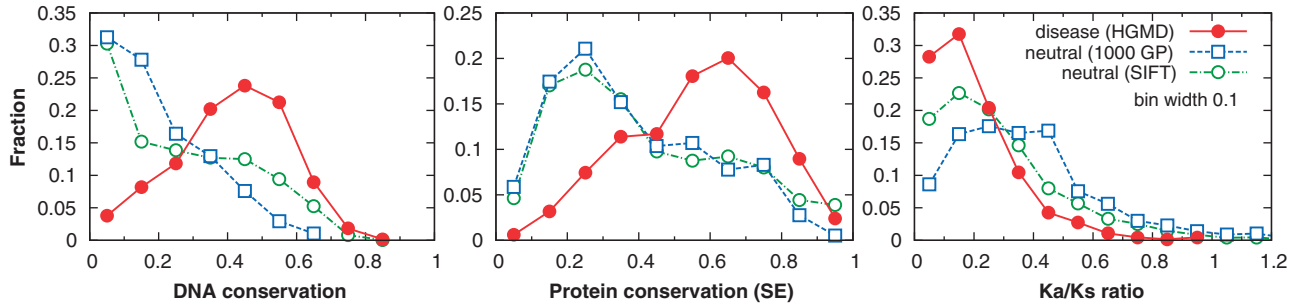

Fig. 1. Distributions of disease-causing FS indels from the HGMD and putatively neutral FS indels from the 1000 GP as well as from the SIFT Indel training dataset (SIFT) for the fraction of conserved DNA bases (phyloP score >2.5) within a window of 40 bases before the indel site and all bases after the indel site, minimum of protein conservation calculated as Shannon entropy (SE) within a window of 5 residues before the indel site and all residues after the indel site, and K_a/K_s mutation ratio for the entire gene. Disease-causing indels tend to be located in regions of highly conserved DNA and protein sequences (here, SE is scaled so that 1 represents the highest conservation), and in genes with a lower number of non-synonymous relative to synonymous substitutions

Table 3. Performance of DDIG-in (FS) (both 10-fold cross-validation and independent test), SIFT Indel, and CADD on the two FS indels datasets

Method	AUC[a]	MCC[a]	Q_2[a]	Se[a]	Sp[a]	PPV[a]	NPV[a]
FS indels dataset							
SIFT Indel	0.62	0.29	0.63	0.92	0.29	0.60	0.77
CADD	0.74	0.38	0.69	0.79	0.59	0.69	0.71
DDIG-in (FS)[b]	0.87 ± 0.006[c]	0.59 ± 0.013[c]	0.79	0.86	0.72	0.78	0.82
HGMD + SIFT FS indels dataset							
SIFT Indel	0.87	0.63	0.80	0.95	0.65	0.73	0.93
CADD	0.72	0.35	0.68	0.75	0.60	0.65	0.71
DDIG-in (FS)[d]	0.83 ± 0.001[c]	0.54 ± 0.004[c]	0.77	0.86	0.67	0.72	0.83

[a]AUC, MCC, Q_2, Se, Sp, PPV and NPV are the area under the ROC curve, Matthews correlation coefficient, binary classification accuracy, sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value, respectively.

[b]Ten-fold cross-validation result (sequence similarity ≤30% for any two folds).

[c]Standard deviation for DDIG-in (FS) over 100 replications of random sampling of the disease-causing FS indels (see Section 2).

[d]Independent test result after training DDIG-in (FS) on the FS indels dataset.

CADD (Kircher *et al.*, 2014), which is a general framework capable of predicting any type of deleterious genetic variations. We used the publicly available web-servers of the two methods to predict FS indels in our dataset. Our method achieved absolute (relative) AUC improvements of 0.25 (40%) and 0.13 (18%) when compared with SIFT Indel and CADD (Table 3), respectively. The respective improvements for MCC were 0.30 (103%) and 0.21 (55%). All three methods performed reasonably well in identifying disease-causing indels (reaching sensitivities of 79–92%), whereas specificity (accuracy on neutral indels) was low for SIFT Indel and CADD (29 and 59%, respectively) but high for DDIG-in (FS) (72%).

3.3 Independent test with HGMD + SIFT FS indels dataset

We employed the HGMD + SIFT FS indels dataset as an independent test set and validation of using the variations from the 1000 GP as our neutral training dataset. DDIG-in (FS) yielded a similar performance in this test, reaching an AUC of 0.83 and MCC of 0.54, compared with 0.87 and 0.59 for the 10-fold cross-validation (Table 3). CADD also retained the same (low) performance (AUC of 0.72, MCC of 0.35), indicating that both CADD and DDIG-in (FS) are robust. However, SIFT Indel had a much better performance on this dataset (MCC of 0.63) than on the dataset with the neutral indels from the 1000 GP (MCC of 0.29), indicating that SIFT Indel cannot generalize as well as DDIG-in (FS) or CADD.

Figure 2b compares the performance of DDIG-in (FS) with the performance of SIFT Indel, CADD, and the top three features in terms of ROC curves for the HGMD + SIFT FS indels dataset. Interestingly, DDIG-in (FS) yields a more significant improvement on this dataset than on the training FS indels dataset when compared with the best single feature (DNA conservation). The improvement was 0.10 (14%) for AUC and 0.15 (38%) for MCC. This indicates that a multi-feature model is more robust than a single-feature classification.

3.4 Independent test with NS variants dataset

To further test the robustness of DDIG-in (FS), we employed NS variants as an independent test for DDIG-in (FS) because a PTC likely has a similar effect as a randomized sequence induced by a FS indel. Also, the distributions of the top three features (see Section 3.1) were highly similar for FS indels (Fig. 1) and NS variants (Supplementary Fig. S1). DDIG-in (FS) yielded an AUC of 0.70 and MCC of 0.31 on the NS variants dataset (Table 4). This is a considerably worse performance than for the FS indels (AUC of 0.87, MCC of 0.59). The performance of CADD also decreased significantly when compared with FS indels. CADD's MCC value was 0.38 for FS indels but only 0.17 for NS variants.

To examine what was responsible for the decrease in prediction performance, we built a dedicated method for NS variants by using the NS variants dataset for feature selection and 10-fold cross-validation as we did for DDIG-in (FS). This procedure resulted in a combination of eight features (Supplementary Table S3) selected for the SVM model, denoted as DDIG-in (NS). DDIG-in (NS) achieved an AUC of 0.72 and MCC of 0.33, which were only marginally higher (both increased by 0.02) than the AUC and MCC of DDIG-in (FS) as shown in Table 4. This result indicates that DDIG-in (FS) is robust and performed well in this stringent test.

One possible reason for the poorer performance on the NS variants dataset is that some of putatively neutral NS variants may be in fact disease-causing. To examine this possibility, we removed the rare neutral NS variants with AFs ≤ 0.05% (2717 removed) and ≤0.1% (3165 removed), resulting in 1144 and 696 neutral variants, respectively. The performance of DDIG-in (NS) then improved significantly with respective AUC improvements from 0.72 to 0.79 and 0.82 (Table 4). DDIG-in (FS) achieved similar improvements.

In fact, DDIG-in (FS) and DDIG-in (NS) yielded exactly the same MCC and AUC when neutral variants with AF $\leq$ 0.1% were removed. Figure 2c depicts the comparison in terms of ROC curves. DDIG-in (FS) continues to outperform CADD for NS variants while having a comparable ROC curve to DDIG-in (NS). This further confirms the robust performance of DDIG-in (FS).

3.5 Independent test by AF

Figure 3 plots the average of $\log_{10}$ disease probability as a function of the average of $\log_{10}$ AF for the 580 neutral FS indels (Fig. 3a) and 3861 neutral NS variants (Fig. 3b) from our two datasets based on the 1000 GP. AFs were grouped into bins so that each bin contained at least 20 variants. However, some bins were larger due to many variants with the same AF value. The disease probabilities were predicted with DDIG-in (FS) for both FS indels and NS variants.

Although there was no significant correlation between individual AF values and predicted disease probabilities (Supplementary Fig. S2), there was a negative correlation between the averaged AF bins and disease probability (Fig. 3) with a Pearson correlation coefficient of −0.33 and −0.65 for FS indels and NS variants, respectively. This means that our method tends to predict higher disease probabilities for lower AFs in the general population. The correct trend obtained here provides another independent validation of DDIG-in (FS).

4 Discussion

This article developed a machine learning method named DDIG-in (FS) that can predict disease probability for both FS indels and NS variants. For FS indels, DDIG-in (FS) has a consistent performance with a MCC of 0.59 in a 10-fold cross-validation and 0.54 in an independent test. Its performance for NS variants matches that of DDIG-in (NS) directly trained for the purpose. Predicted disease probabilities for neutral variants from the 1000 GP are supported by a negative correlation with the average AFs of these variants.

In developing our method, we performed feature selection and classifier obtainment using the same FS indels dataset (Table 1) because it is the only human dataset available. Because this practice may result in over-training, several techniques were utilized to minimize such a possibility. To avoid over-training in a particular group of genes, we used a sequence identity threshold (30%) to remove redundant sequences in the disease-causing FS indels dataset. For the 10-fold cross-validation, we ensured that no two folds shared sequences with a pairwise identity >30%. A similar approach was

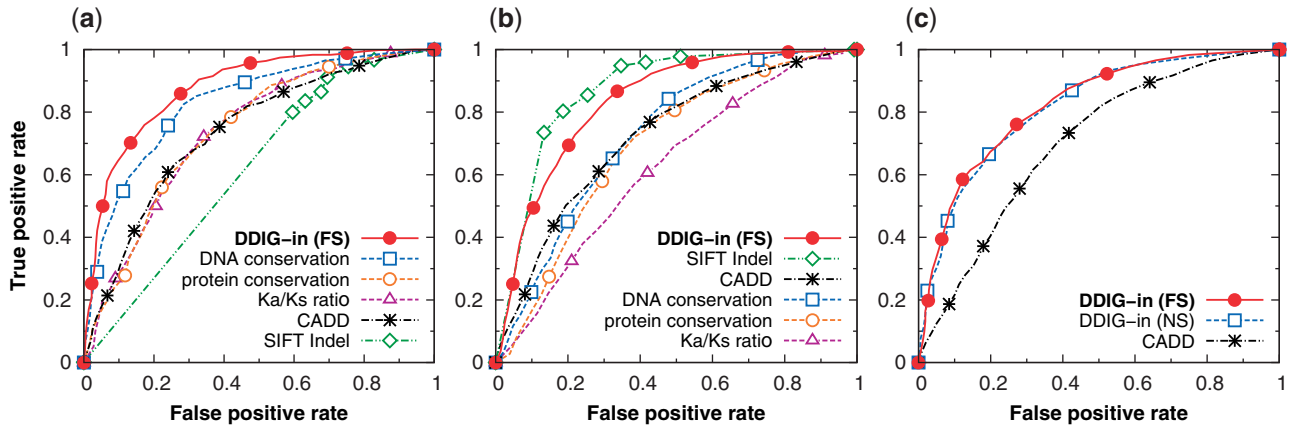

Fig. 2. ROC curves of DDIG-in (FS) as compared with those of SIFT Indel, CADD, and the top three features [DNA conservation, protein conservation (SE), and K_a/K_s ratio]. **(a)** FS indels dataset: DDIG-in (FS) is the 10-fold cross-validated result. **(b)** HGMD + SIFT FS indels dataset: DDIG-in (FS) is the independent test result. **(c)** NS variants dataset: DDIG-in (FS) is the independent test result, whereas DDIG-in (NS) is from 10-fold cross-validation; for clarity, we only showed the results of DDIG-in (FS), DDIG-in (NS), and CADD after neutral variants with AF $\leq$ 0.1% were removed. Essentially the same performance of DDIG-in (FS) and DDIG-in (NS) for NS variants demonstrates the robustness of DDIG-in (FS)

Table 4. Performance of DDIG-in (FS) (independent test), DDIG-in (NS) (10-fold cross-validation), and CADD for discriminating NS variants

Method	Neutral variants[a]	AUC[b]	MCC[b]	Q_2[b]	Se[b]	Sp[b]	PPV[b]	NPV[b]
CADD	AF $\geq$ 0.00%	0.60	0.17	0.57	0.84	0.30	0.55	0.65
CADD	AF > 0.05%	0.67	0.26	0.74	0.84	0.42	0.83	0.44
CADD	AF > 0.10%	0.70	0.28	0.78	0.84	0.47	0.90	0.35
DDIG-in (FS)	AF $\geq$ 0.00%	0.70 ± 0.001[c]	0.31 ± 0.002[c]	0.64	0.86	0.41	0.59	0.75
DDIG-in (FS)	AF > 0.05%	0.79 ± 0.001[c]	0.42 ± 0.002[c]	0.80	0.88	0.53	0.86	0.57
DDIG-in (FS)	AF > 0.10%	0.82 ± 0.001[c]	0.43 ± 0.002[c]	0.85	0.92	0.50	0.91	0.53
DDIG-in (NS)	AF $\geq$ 0.00%	0.72 ± 0.003[c]	0.33 ± 0.003[c]	0.66	0.72	0.60	0.64	0.69
DDIG-in (NS)	AF > 0.05%	0.79 ± 0.002[c]	0.43 ± 0.006[c]	0.81	0.91	0.47	0.85	0.62
DDIG-in (NS)	AF > 0.10%	0.82 ± 0.002[c]	0.43 ± 0.007[c]	0.86	0.92	0.50	0.91	0.53

[a]Neutral variants in the dataset were filtered by AF at three thresholds: AF $\geq$ 0% (no filtering), AF > 0.05%, and AF > 0.1%.

[b]AUC, MCC, Q_2, Se, Sp, PPV and NPV are the area under the ROC curve, Matthews correlation coefficient, binary classification accuracy, sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value, respectively.

[c]Standard deviation for DDIG-in (FS) (independent test) and DDIG-in (NS) (10-fold cross-validation) over 100 replications of random sampling of the disease-causing NS variants.

used previously for methods for the prediction of damaging missense variants (Adzhubei *et al.*, 2010; Li *et al.*, 2009) or the prediction of mutation-induced stability changes (Folkman *et al.*, 2014a,b). We replicated our experiments 100 times with randomly generated folds while maintaining the ratio of the disease-causing and neutral variants. This strict cross-validation coupled with the sequential forward floating feature selection (Pudil *et al.*, 1994) led to a robust performance in two fully independent tests, which in turn confirmed that DDIG-in (FS) was not over-trained (Tables 3 and 4).

Our method differs from two other available methods, SIFT Indel (Hu and Ng, 2012) and CADD (Kircher *et al.*, 2014), in training. CADD was trained with disease-causing variants from a genome-wide simulation of *de novo* germline mutations and neutral variants between human and inferred human-chimpanzee common ancestral genomes. SIFT Indel was trained with disease-causing FS indels from the HGMD and neutral variants between human and orthologous animal proteins. By contrast, both disease-causing and neutral indels for training DDIG-in (FS) were from known human variations (HGMD and 1000 GP, respectively).

Our new method also differs from previous techniques by utilizing features from predicted structural properties of proteins. Nonetheless, the three most prominent features in our method were all related to sequence conservation (Table 2). In agreement with related work [SIFT Indel (Hu and Ng, 2012)], DNA conservation was the single-most important feature, while those features based on predicted structural properties of proteins were significantly less discriminative. By comparison, the most discriminative feature for NFS indels (Zhao *et al.*, 2013) was the disruption of structured regions predicted by protein disorder predictor SPINE-D (Zhang *et al.*, 2012). This highlights the difference between local sequence disruption induced by NFS indels and the global sequence disruption induced by FS indels and NS variants. Indeed, the direct application of DDIG-in (FS) to a NFS indels dataset results in a significantly poorer performance [MCC of 0.42 compared with 0.68 by DDIG-in (NFS) (Zhao *et al.*, 2013)].

Unlike previously developed methods, DDIG-in (FS) was subjected to rigorous independent testing by applying it directly to NS variants. Here, we assumed the equivalence between FS indels and NS variants because FS indels effectively render the protein sequence meaningless after the indel site, and probably have the same effect as truncation of the protein induced by a NS variant. This equivalence was supported by similar distributions of three predictive features (DNA conservation, protein conservation, and K_a/K_s ratio) for NS variants and FS indels (Supplementary Fig. S1 and Fig. 1, respectively). The equivalence is further supported by the fact that DDIG-in (NS), directly optimized and trained for NS variants, was only marginally better (AUC of 0.72 compared with 0.70) than the independent test performance of our 'main' method, DDIG-in (FS), which was designed solely using a dataset of FS indels (see Section 3.4). The performance of DDIG-in (FS) was lower for NS variants as compared with FS indels, but the possible difference between these two distinct types of variants is unlikely to be the main cause. We found that the neutral NS variants dataset comprised a significant proportion of extremely rare variants with AF $\leq 0.1\%$. By comparison, AFs for all neutral FS indels were $\geq 0.3\%$. When we tested DDIG-in (FS) only on neutral NS variants with AF $\geq 0.3\%$, the prediction performance was comparable to the performance on FS indels (AUC of 0.86 and 0.87, respectively). Thus, the main reason for a lower performance of DDIG-in (FS) on NS variants was the presence of false negatives in the neutral NS variants dataset.

Another means of independent testing is by the use of AF, the occurrence frequency of a particular variant in a given population. AF results from multiple factors arising from the complicated interactions between human beings and their environment. Many rare alleles are both population-specific and of functional significance (Marth *et al.*, 2011), whereas frequent alleles (for late-onset diseases, in particular) are not necessarily benign. Thus, the fitness of the allele with respect to its associated biological function is likely an underlying trend that appears only after averaging (removing noise from other factors) (Hu and Ng, 2013; Zhao *et al.*, 2013). This correlation with the average AF values but not with the individual AF values is illustrated in Figure 3 and Supplementary Figure S2, respectively.

In this work, we assumed that FS indels and NS variants identified by the 1000 GP are neutral. Although this assumption is not unreasonable, there might be false negatives (disease-causing variants labelled as neutral) in our datasets. Indeed, rare neutral NS variants (AF $\leq 0.1\%$) have led to poorer performance of CADD, DDIG-in (FS), and DDIG-in (NS) (Table 4). Interestingly, we found that DDIG-in (NS) trained and evaluated only on neutral variants with AF $> 0.1\%$ yields an AUC of 0.76, significantly worse than 0.82 given by DDIG-in (NS) trained with all neutral variants. Thus, the benefit resulting from a larger number of neutral variants outweighs the potential for false negatives for some rare variants. Similar behaviour was observed for NFS indels (Zhao *et al.*, 2013). Utilizing variants from the 1000 GP as our neutral dataset is further supported by the consistent performance of DDIG-in (FS) when these neutral variants were replaced by variants derived from animal orthologs from the SIFT Indel training dataset. Moreover, the distributions of sequence conservation features for neutral FS indels from the 1000 GP are similar to those from the SIFT Indel training dataset (Fig. 1).

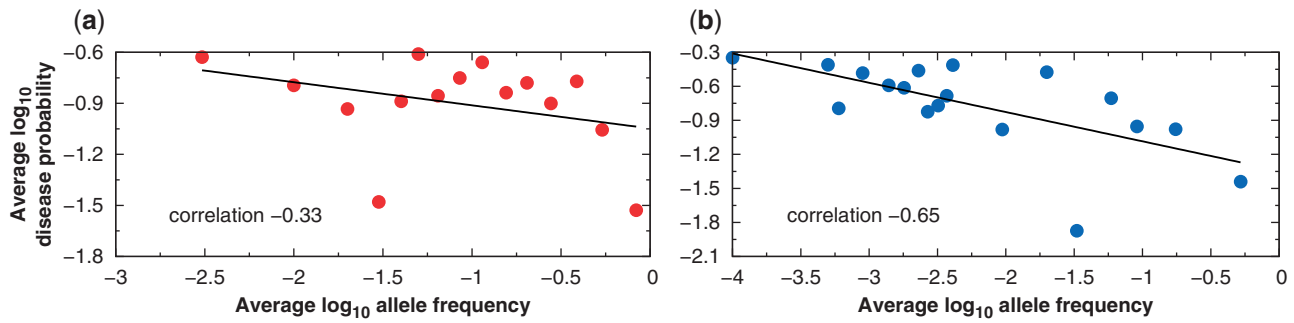

Fig. 3. Average $\log_{10}$ disease probability as a function of the average $\log_{10}$ AF for the neutral FS indels and NS variants from our two datasets based on the 1000 GP. All disease probabilities were predicted with DDIG-in (FS), the average being calculated for each AF bin containing at least 20 variants. The black line is the linear regression fit. (a) FS indels dataset and (b) NS variants dataset

Acknowledgement

The authors thank Jing Hu for sharing the SIFT Indel neutral dataset.

Funding

This work was supported in part by National Health and Medical Research Council (1059775) of Australia to Y.Z. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, and the Australian Research Council through the ICT Centre of Excellence program. The authors also gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster 'Gowonda' to complete this research. This project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al*. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods,* **7**, 248–249.

Altschul,S. *et al*. (1990). Basic local alignment search tool. *J. Mol. Biol.,* **215**, 403–410.

Altschul,S. *et al*. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.,* **25**, 3389.

Ball,E.V. *et al*. (2005). Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.,* **26**, 205–213.

Bendl,J. *et al*. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.,* **10**, e1003440.

Bermejo-Das-Neves,C. *et al* (2014). A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). *BMC Bioinformatics,* **15**, 111.

Capra,J.A. and Singh,M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics,* **23**, 1875–1882.

Choi,Y. *et al*. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One,* **7**, e46688.

Cortes,C. and Vapnik,V. (1995). Support-vector networks. *Mach. Learn.,* **20**, 273–297.

Faraggi,E. *et al*. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure,* **17**, 1515–1527.

Faraggi,E. *et al.* (2011). SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.,* **33**, 259–267.

Flicek,P. *et al.* (2013). Ensembl 2014. *Nucleic Acids Res.,* **42**, D749–D755.

Folkman,L. *et al* (2014a). Feature-based multiple models improve classification of mutation-induced stability changes. *BMC Genomics,* **15**(Suppl. 4), S6.

Folkman,L. *et al.* (2014b). Towards sequence-based prediction of mutation-induced stability changes in unseen non-homologous proteins. *BMC Genomics,* **15**(Suppl. 1), S4.

Hu,J. and Ng,P.C. (2012). Predicting the effects of frameshifting indels. *Genome Biol.,* **13**, R9.

Hu,J. and Ng,P.C. (2013). SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One,* **8**, e77940.

Hurst,L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.,* **18**, 486–487.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.,* **32**(Suppl. 1), D493–D496.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.,* **46**, 310–315.

Li,B. *et al.* (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics,* **25**, 2744–2750.

Marth,G.T. *et al.* (2011). The functional spectrum of low-frequency coding variation. *Genome Biol.,* **12**, R84.

McVean,G.A. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature,* **467**, 1061–1073.

Mills,R.E. *et al.* (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.,* **21**, 830–839.

Mort,M. *et al.* (2008). A meta-analysis of nonsense mutations causing human genetic disease. *Hum. Mutat.,* **29**, 1037–1047.

Mort,M. *et al.* (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.,* **15**, R19.

Nagy,E. and Maquat,L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.,* **23**, 198–199.

Ng,P. and Henikoff,S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.,* **11**, 863–874.

Pollard,K.S. *et al.* (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.,* **20**, 110–121.

Pudil,P. *et al.* (1994). Floating search methods in feature selection. *Pattern Recogn. Lett.,* **15**, 1119–1125.

Remmert,M. *et al.* (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods,* **9**, 173–175.

Scholkopf,B. *et al.* (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.,* **13**, 1443–1471.

Stenson,P.D. *et al.* (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.,* **133**, 1–9.

Thusberg,J. *et al.* (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.,* **32**, 358–368.

van Hoof,A. *et al.* (2002). Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science,* **295**, 2262–2264.

Zhang,T. *et al*. (2012). SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *J. Biomol. Struct. Dyn.,* **29**, 799–813.

Zhang,X. *et al.* (2014). Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation. *Hum. Mol. Genet.,* **23**, 3024–3034.

Zhao,H. *et al.* (2013). DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol.,* **14**, R23.

Zia,A. and Moses,A.M. (2011). Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics,* **12**, 299.