BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2181–2186
doi:10.1093/bioinformatics/btr365

Genome analysis Advance Access publication June 17, 2011

KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels

L. F. Stead[1,†], I. C. Wood[2] and D. R. Westhead[1,*]

[1]Institute of Molecular and Cellular Biology, Faculty of Biological Sciences and [2]Institute of Membrane and Systems Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Non-synonymous single nucleotide polymorphisms (nsSNPs) in voltage-gated potassium (Kv) channels cause diseases with potentially fatal consequences in seemingly healthy individuals. Identifying disease-causing genetic variation will aid presymptomatic diagnosis and treatment of such disorders. NsSNP-effect predictors are hypothesized to perform best when developed for specific gene families. We, thus, created KvSNP: a method that assigns a disease-causing probability to Kv-channel nsSNPs.

Results: KvSNP outperforms popular non gene-family-specific methods (SNPs&GO, SIFT and Polyphen) in predicting the disease potential of Kv-channel variants, according to all tested metrics (accuracy, Matthews correlation coefficient and area under receiver operator characteristic curve). Most significantly, it increases the separation of the median predicted disease probabilities between benign and disease-causing SNPs by 26% on the next-best competitor. KvSNP has ranked 172 uncharacterized Kv-channel nsSNPs by disease-causing probability.

Availability and Implementation: KvSNP, a WEKA implementation is available at www.bioinformatics.leeds.ac.uk/KvDB/KvSNP.html.

Contact: d.r.westhead@leeds.ac.uk

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

Received on April 12, 2011; revised on June 4, 2011; accepted on June 8, 2011

1 INTRODUCTION

One in every 300 bases of human DNA shows variation between individuals. These sites are called single nucleotide polymorphisms (SNPs; The International HapMap Consortium, 2003). Distinguishing benign SNPs (bSNPs) from those that cause disease (dcSNPs) is a key premise for personalized medicine, but with a total of 10 million polymorphic loci within the genome it is no trivial feat. Efforts are often focused on identifying SNPs responsible for specific diseases (Kapa *et al.*, 2009), or all dcSNPs within certain genomic regions (The ENCODE Consortium, 2007) or genes (Splawski *et al.*, 2000); for example, identifying dcSNPs within the voltage-gated potassium (Kv) channel genes, where genetic variation has been shown to cause, or be associated with, cardiac arrhythmogenesis, for which the first presenting symptom can be sudden cardiac death. Genetic screening of a Long QT Syndrome (LQTS) cohort of 262 patients revealed that 59% had at least 1 minor allele in a Kv-channel gene (Splawski *et al.*, 2000). An additional concern with arrhythmogenesis is that individuals with certain genetic backgrounds (including Kv-channel genotypes) can acquire it when administered with certain drugs (Anantharam *et al.*, 2003). Furthermore, SNPs in Kv-channel genes also cause, or predispose individuals to, various neurological disorders including epilepsy (Kaneko *et al.*, 2002), neuromyotonia (Poujois *et al.*, 2006) and episodic ataxia (Scheffer *et al.*, 1998), as well as other diseases, such as deafness (Kubisch *et al.*, 1999). While not directly fatal, these disorders would still benefit from pre-symptomatic diagnosis and treatment.

Association studies have previously been used to identify Kv-channel dcSNPs (Kapa *et al.*, 2009), but there is concern that some variants may be incorrectly assigned because study cohorts were not large enough to infer statistical significance (Ackerman *et al.*, 2003; Kääb and Schulze-Bahr, 2005). Results are also confounded by multiple testing errors, and the fact that compound SNPs are likely to cause, or contribute to, disease (Brookes, 1999). Variable phenotype penetrance and severity is observed in arrhythmia patients harboring the same Kv-channel SNP (Scicluna *et al.*, 2008), suggesting that phenotypic modifier genes and/or variants exist (Schwartz *et al.*, 2003). Ranking SNPs by disease-causing probability would (i) prioritize variants for further examination and (ii) aid identification of phenotype-modifying variants. This is made even more relevant by the current sequencing revolution, where advances in high-throughput technology have facilitated genotyping to an unprecedented depth at affordable costs; resulting in a plethora of data that some are finding hard to analyse fast enough (Editorial, 2008).

Computational methods to predict whether nsSNPs will cause disease were originally based on empirical rules (Ng and Henikoff, 2001; Ramensky *et al.*, 2002) but now tend to use machine learning approaches (Calabrese *et al.*, 2009; Jiang *et al.*, 2007; Krishnan and Westhead, 2003; Yue and Moult, 2006). Variants of known consequence are used to train the machine learning predictor, which finds patterns within the annotated nsSNP attributes that best classify the data into in their constituent groups: dcSNPs and bSNPs. Once trained, the predictor can be applied to uncharacterized variants, emitting the likelihood that they are disease causing. This approach has resulted in prediction accuracies of up to 82% (Calabrese *et al.*, 2009), but there is some evidence that performance can be improved if the machine learning model is built for a specific gene

*To whom correspondence should be addressed.
†Present address: Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, Leeds LS9 7TF, UK.

family. This is hypothesized to be due to the gene family specific weighting of variant attributes within the model, and the ability to include attributes that are biologically relevant to those genes and/or protein products (Torkamani and Schork, 2007). To test this, we built an nsSNP prediction method specifically for Kv-channel genes, KvSNP, and optimized it with respect to the learning features and machine learning method. KvSNP outperforms the leading genome-wide SNP effect prediction method with increased separation of bSNPs and dcSNPs in probability space. The latter is particularly pertinent as the trust placed in the accuracy of disease annotation in training datasets for SNP-effect predictors is a point of concern (Care *et al.*, 2007), especially for complex diseases where multiple SNPs are expected to contribute to disease phenotype, thus confounding, or removing, the ability to classify individual variants as disease causing or benign (Ackerman *et al.*, 2003).

Genetic characterization of Kv-channels will aid presymptomatic diagnosis and treatment of diseases such as cardiac arrhythmia. To this end, we have used KvSNP to predict the disease-causing ability of 172 uncharacterized nsSNPs and present case studies regarding those predictions.

2 METHODS

2.1 The Dataset

Data on characterized nsSNPs were collated from Uniprot 13 (The UniProt Consortium, 2009), the Human Gene Mutation Database: HGMD Professional 2009.2 (Stenson *et al.*, 2009), MutDB (Singh *et al.*, 2008) and manual extraction from research papers (Supplementary Table S1). A dcSNP annotation (irrespective of the disease) was assigned if concluded by (i) the genotype–phenotype database curators and/or (ii) the authors of the research papers. Conflicting annotations were resolved by assigning that given by the largest and most statistically robust study. Truncation mutations were not included as all were annotated as disease causing, hence the machine learning method would be unable to learn from them. This resulted in 1009 dcSNPs and 113 bSNPs. To increase the number of bSNPs, we added to them by identifying single nucleotide differences between human and homologous mammalian Kv genes (>90% sequence identity) that result in a change in protein sequence, the assumption being that fixed evolutionary changes are neutral. Such 'divergent' datasets have been used in similar ways previously (Care *et al.*, 2007; Sunyaev *et al.*, 2001; Yue and Moult, 2006) and independent Kv-channel research into disease susceptibility has shown that a human nsSNP that is present in a homologue is unlikely to be disease causing (Jackson and Accili, 2008). Our final dataset consisted of 1009 dcSNPs and 782 bSNPs, of which 100 dcSNPs and 76 bSNPs (∼10%) were removed to form a blind test dataset. This data removal was performed at random but multiple variant protein products resulting from a single genetic variation (owing to alternative splicing) were filtered to ensure they were wholly contained in either the training or test data. The remaining training dataset consisted of 706 bSNPs and 909 dcSNPs.

2.2 Learning features

A total of 14 nsSNP attributes were used as learning features; summarized in Table 1. The wild-type and mutant residue identity are included as a basic descriptor of the nsSNP. Evolutionary conservation of the variant position has been shown to be predictive (Care *et al.*, 2007; Jiang *et al.*, 2006), hence conservation scores were assigned using Rate4Site (Mayrose *et al.*, 2004), a top-performing method (Capra and Singh, 2007; Dukka Bahadur and Livesay, 2008). Hidden Markov Model (HMM) sequence profiles, created using HMMER (Eddy, 1998), were used to assign two additional features: change in HMM score and change in HMM *E*-value, which quantify how an nsSNP alters a sequence's fit to the modelled profile (Clifford *et al.*, 2004).

Table 1. The learning features used to annotate each protein sequence variant in our dataset

Learning feature	Description	Label
Wild-type residue	Identity of the wild-type amino acid residue	wt
Mutant residue	Identity of the variant amino acid residue	mut
Conservation score	A multiple sequence alignment-based metric	cons
Subfamily membership	Classification of protein sequences due to homology	subfam
Topological location	The protein region where the variant occurs (Fig. 1)	loc
Change in hydrophobicity	Between the wild-type and variant residue	Hphob
Change in molecular mass	Between the wild-type and variant residue	mass
Change in isolelectric point	Between the wild-type and variant residue	pI
Predicted secondary structure	Predicted from protein sequence	SS
Predicted solvent accessibility	Predicted from protein sequence	rsa
Change in HMM sequence profile score	$Score_{var}$-$Score_{wt}$/ $Score_{var}$. Score is that resulting from querying the variant (var) or wild-type (wt) sequence against a subfamily specific HMM	HMM
Change in HMM sequence *E*-value	$Log_{10}(E_{var}/E_{wt})$. *E* is the *E*-value when the variant (var) or wild-type (wt) sequence is queried against a subfamily-specific HMM	logRE
Predicted change in stability	A potential energy function-based metric using structural models of the proteins	dE
Predicted solvent accessibility	Predicted using proteins structural models	solv.acc

The label indicates the short hand used to refer to each feature in tables and figures throughout this article.

Conservation scoring and HMM creation require multiple sequence alignments (MSAs) which were created as we detailed previously (Stead *et al.*, 2010). The majority of Kv-channel LQTS-causing variants reside within the transmembrane segments (Ackerman *et al.*, 2003) and our previous research has revealed significant associations between disease-causing probability and location of nsSNPs, and hence topological location (Fig. 1) was included, assigned as per Stead *et al.* (2010).

Structure-based learning features can improve SNP-effect predictors (Krishnan and Westhead, 2003), hence we built homology models in Modeller (Sali and Blundell, 1993), using the only resolved high-resolution Kv-channel structure (2R9R: Long *et al.*, 2007) and predict how each nsSNP changes stability and solvent accessibility using PoPMuSiC2.0, which is both fast and accurate (Dehouck *et al.*, 2009).

Changes in physiochemical property at the variant locus may also disrupt to local interactions, hence changes in mass, hydrophobicity and isoelectric point were extracted from Biro (2006). Secondary structure and solvent accessibility were also predicted from sequence alone, using SABLE (Adamczak *et al.*, 2005), a popular programme for protein structure analysis (Gromiha *et al.*, 2010). Subfamily membership was included owing to its predictive power when included in a kinase-specific SNP effect prediction method (Torkamani and Schork, 2007).

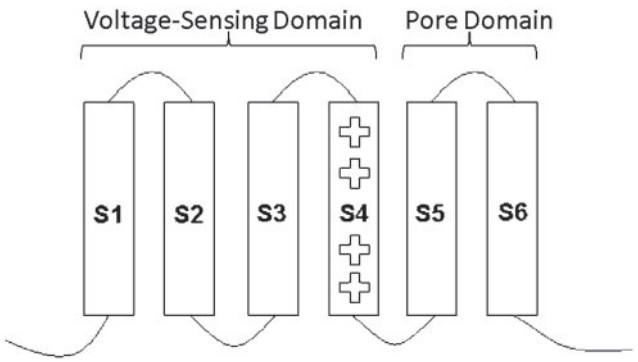

Fig. 1. Kv-channel core topology. Transmembrane segments (S1-6) create two functional domains: voltage-sensing (S1-4) and pore domain (S5-6).

2.3 The machine learning method

Several machine learning methods were tested and implemented in the Weka (Hall *et al.*, 2009):

- Fast Random Forest (FRF), an implementation of the random forest, RF (Breiman, 2001), classifier that uses the REPTree algorithm for decision trees (DTs);
- J48, a DT based on the C4.5 (Quinlan, 1993) that uses information gain for feature selection at each node, and prunes using subtree raising;
- a support vector machine (SVM) that uses sequential minimal optimization;
- a naïve Bayes classifier (John and Langley, 1995); and
- an artificial neural network (ANN). A multilayer perceptron was used with sigmoidal transfer function.

Each machine learning methods assigns queries with a disease-causing probability [p(dcSNP)]. DTs do this based on the weighting of classes at each terminal node while RFs amalgamate the results from all constituent DTs. The SVM assigns probabilities using logistic regression, and the naïve Bayes classifier results are based on class frequencies in the training set. The results from the ANN are based on the weighted output of flow across the network with a sigmoidal transfer function applied. A nsSNP is classified as a dcSNP if $P \geq 0.5$ for all these methods.

2.4 Performance metrics

Ten-fold, stratified, cross-validation (CV) was used to assess the performance of each machine learning method, revealing counts of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). We used these counts to ascertain accuracy [Equation (1)], Matthews correlation coefficient [MCC: Equation (2)] and area under the receiver operating characteristic (ROC) curve (AUC) for each method.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \quad (1)$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (2)$$

3 IMPLEMENTATION: KvSNP

3.1 Comparison with other methods

The blind test data (which was not used in the CV above) was used to assess KvSNP's performance and compare it with that of the existing genome-wide nsSNP-effect predictors:

- SNPs&GO: an SVM-based method that includes gene ontology (GO) annotations in its learning features (Calabrese *et al.*, 2009);
- Sorting intolerant from tolerant (SIFT): predictions based on conservation scores from MSAs (Ng and Henikoff, 2003).
- PolyPhen (polymorphism phenotyping): predictions made using empirical rules (Ramensky *et al.*, 2002).

Results were either attained from web servers (SNPs&GO and PolyPhen) or using locally downloaded copies of the software (SIFT). The blind test data may have been used in training existing methods, but this possible advantage to their performance cannot be avoided easily. To allow significance testing and error estimation, the blind test data were randomly split into 10 stratified folds and metrics were calculated for each fold before being averaged.

3.2 Making predictions on uncharacterized nsSNPs

Whilst sourcing our dataset, we also collated information on 172 Kv-channel nsSNPs for which an effect is unknown, using dbSNP 129 (Sherry *et al.*, 2001) as an additional resource. We predicted the effect of these using the KvSNP server.

4 RESULTS

4.1 A set of 5 learning features performs best

To determine which learning features to use in our method, we applied a best-first search algorithm (Pearl, 1984) where each feature is evaluated and the best, based on correlation-based feature selection (Hall, 2000), is expanded upon to form the next 'node' in the search. This process resulted in five learning features: conservation score (cons), change in hydrophobicity (Hphob), change in structural stability (dE), change in HMM score (HMM) and subfamily membership (subfam).

4.2 The FRF classifier works best

We optimized the parameters for each machine learning classifier, using the training data and set of five best learning features, by testing a range of values and selecting those combinations at which the model attained maximum MCC and accuracy (data not shown). This resulted in:

- a DT with a pruning confidence threshold of 0.27, minimum of 1 instance per leaf and that output leaf conditional probabilities with no Laplace smoothing;
- an FRF created from 14 DTs with 3 features selected per node;
- a polynomial-kernel SVM with soft margin constant of 3 and class probability estimates equated using logistic regression;
- a naive Bayes classifier that modeled continuous variables using supervised discretization; and
- an ANN with a learning rate of 0.1 and momentum of 0.2.

Performance metrics for the optimized classifiers are given in Figure 2 and Table 2. All were normally distributed (Shapiro–Wilk: $P > 0.5$). The *t*-tests showed that FRF had a significantly increased AUC compared with all other methods, and a significantly increased MCC and accuracy compared with all other methods except DT ($P < 0.05$). We, thus, chose FRF as the machine learning method to creating KvSNP.

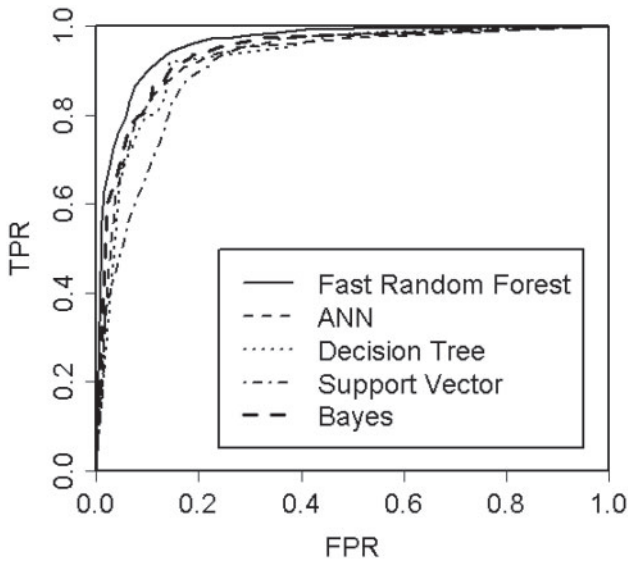

Fig. 2. ROC plots for different machine learning methods. TPR, true positive rate; FPR, false positive rate.

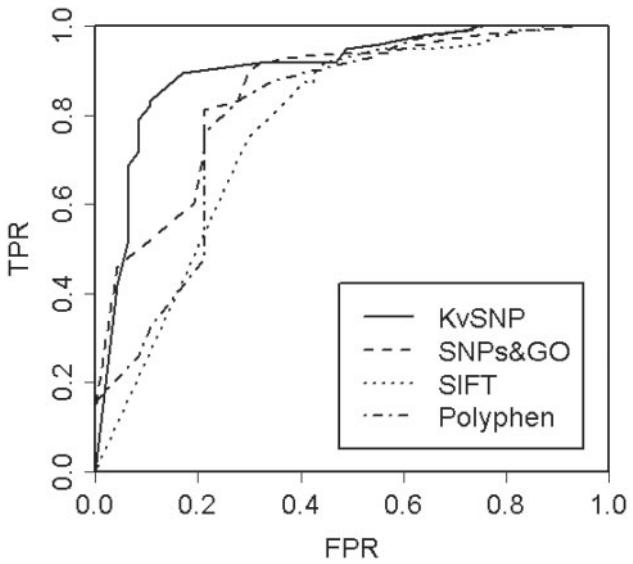

Fig. 3. ROC plots for different nsSNP-effect predictors.

Table 2. Classifier performance ± SEM

	DT	FRF	Bayes	SVM	ANN
Accuracy	0.89 ± 0.01	0.90 ± 0.01	0.87 ± 0.01	0.86 ± 0.01	0.87 ± 0.01
MCC	0.77 ± 0.01	0.80 ± 0.02	0.75 ± 0.01	0.71 ± 0.02	0.74 ± 0.02
AUC	0.92 ± 0.01	0.96 ± 0.01	0.95 ± 0.01	0.91 ± 0.01	0.93 ± 0.01

Table 3. Different nsSNP-effect predictor performance ± SEM

	KvSNP	SNPs&GO	SIFT	PolyPhen
Accuracy	0.87 ± 0.03	0.83 ± 0.03	0.73 ± 0.04	0.72 ± 0.05
MCC	0.70 ± 0.05	0.62 ± 0.07	0.42 ± 0.07	0.44 ± 0.07
AUC	0.92 ± 0.02	0.85 ± 0.04	0.79 ± 0.04	0.81 ± 0.04

4.3 KvSNP outperforms alternative methods

KvSNP was then compared with several other genome-wide nsSNP-effect predictors using the blind test data. The metrics were normally distributed (Shapiro–Wilk test: $P > 0.05$). Results (Fig. 3 and Table 3) show that, in comparison to the next-best performing predictor, KvSNP has significantly (t-test: P-values given below) greater accuracy (0.87 ± 0.03 versus 0.83 ± 0.03: $P = 0.16$), MCC (0.70 ± 0.05 versus 0.62 ± 0.07: $P = 0.09$) and AUC (0.92 ± 0.02 versus 0.85 ± 0.04: $P = 0.07$).

To be effective, prediction methods should clearly distinguish dcSNPs from bSNPs (Jiang *et al.*, 2007) by assigning significantly different disease-causing probabilities [p(dcSNP)]. The distributions of p(dcSNP), per class, are plotted in Figure 4 and are significantly different for each method (t-test: $P < 0.05$). However, Figure 4 clearly shows that KvSNP produces a significantly greater (t-test: $P < 0.05$ except against SNP&GO where $P = 0.08$) separation of classes with a difference in medians (±SEM) of 0.63 ± 0.06 for KvSNP, 0.50 ± 0.06 for SNPs&GO, 0.36 ± 0.06 for SIFT and 0.23 ± 0.03 for PolyPhen, i.e. increase in separation of classes of 26% over the next-best method.

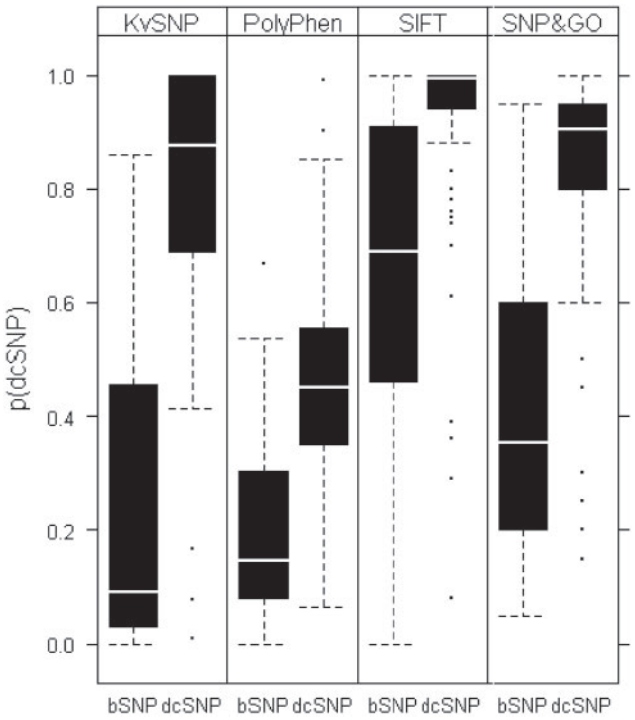

Fig. 4. Disease-causing probability distributions [p(dcSNP)] for different nsSNP-effect predictors. The horizontal white line denotes the median.

4.4 KvSNP predictions on uncharacterized nsSNPs

We applied KvSNP to the uncharacterized variants and ranked the results in descending order of disease-causing probability (Supplementary Table S2). An additional literature search, in an attempt to verify our predictions, did not reveal that any of these variants have been statistically classified since we collated our dataset.

However, there were several findings worthy of note.

KCNQ1 H105L (*NM_000218.2:c.381C>A*): this was identified in a single individual during a gene mutation screening of sudden infant death syndrome (SIDS) victims (Wedekind *et al.*, 2006). SIDS has been linked to cardiac arrhythmia (Makielski, 2006). Electrophysiological characterization did not reveal any alteration in channel function *in vitro*, but KvSNP predicted this to be a disease causing SNP with a probability [p(dcSNP)] of 1.0. We believe this variant should be investigated further *in vivo* to ascertain if it should be part of clinical genetic screening.

KCNA5 G182R (NM_002234.2:c.544G>A) and E211D (NM_002234.2:c. 633G>C): both observed in heterozygous form in patients (2 per variant) suffering idiopathic pulmonary arterial hypertension (IPAH) (Remillard *et al.*, 2007). The cohort numbers used in this study are not sufficient to statistically assign the nsSNPs as disease causing, hence electrophysiological assays were carried out. The assays revealed that G182R significantly changes certain channel properties (inactivation kinetics and protein levels) with respect to wild-type, whereas E211D does not (Burg *et al.*, 2010). KvSNP predicts that *KCNA5* G182R is disease causing [p(dcSNP) = 1.0], whereas E211D is not [p(dcSNP) = 0]. G182R should, thus, be investigated further *in vivo*.

KCNH2 A1116V (*NM_000238.2:c.3347C>T*): implicated as a latent LQTS causal variant, benign unless *KCNH2* K897T (NM_000238.2:c.2690A>C) is also present (Crotti *et al.*, 2005). KvSNP assigns a p(dcSNP) of 0.26, highlighting the benefit of assigning a probability, as opposed to just a class; variants can be ranked in the order of disease-causing susceptibility. *KCNH2* A1116V is predicted to be benign ($0 \leq p(\text{dcSNP}) < 0.5$) but variants ranked highly among the bSNPs are hypothesized to be those most likely to act as genetic modifiers. A1116V is in the top 25% of the uncharacterized variants predicted to be benign (Supplementary Table S2), which concurs with our hypothesis. We could not make an unbiased prediction for *KCNH2* K897T as this variant was included in our training dataset.

5 DISCUSSION

Genetic characterization of Kv-channels is required to identify SNPs that cause cardiac arrhythmia, a disease with potentially fatal consequences for seemingly healthy individuals and other serious diseases. We have created KvSNP: a machine learning classifier that is applied to Kv-channel nsSNPs to predict the likelihood that the variant causes disease. Alternative types of mutation are also known to cause disease in Kv-channels, and these are reported on KvDB (Stead *et al.*, 2010), but too few are characterized to facilitate a machine learning approach for their predicted effect. We found that five learning features optimized KvSNP's performance. Two of these (cons and HMM) are evolutionary descriptors, concurring with previous findings that evolutionary metrics increase predictive power (Care *et al.*, 2007; Jiang *et al.*, 2006), and one is subfamily membership, a feature also included upon creation of a specific nsSNP-effect predictor for the kinase gene family (Torkamani and Schork, 2007). This indicates that gene-family-specific features are important in creating prediction methods, and their inclusion is likely to be part of the reason that KvSNP outperformed alternative genome-wide nsSNP-effect predictors. It should also be noted that we could not verify whether the machine learning based generic nsSNP-effect predictor, SNPs&GO, was trained using any of the blind test data: a fact that would give it an advantage over our method.

Misclassification of SNPs in the training data, a concern with complex disease phenotypes, creates noise in the training dataset, making the machine learning task of discriminating and separating the classes more difficult. We show that KvSNP is best able to separate bSNPs and dcSNPs in probability space, compared with genome-wide predictors, indicating its robustness to noise in the training data, and resulting in disease-causing probabilities that are easier for users to interpret. Overall, this concurs with the 2007 findings of Torkamani and Schork: that gene-family-specific nsSNP-effect predictors are more accurate than genome-wide methods, justifying their creation at a time when the number of novel nsSNPs, in need of characterization, is set to rapidly increase.

6 CONCLUSION

KvSNP is better at discerning the outcome of Kv-channel nsSNPs than genome-wide methods, and is able to discriminate between classes of nsSNPs more effectively. The inclusion of evolutionary descriptors in two of five learning features used highlights the predictive power of such metrics. We have used KvSNP to make predictions on, and thus rank 172 uncharacterized Kv-channel gene nsSNPs to prioritize them for future genetic association studies and identify potential genetic modifiers of phenotype. The KvSNP software and our test and training data are available for download at www.bioinformatics.leeds.ac.uk/KvDB/KvSNP.html.

ACKNOWLEDGEMENTS

We acknowledge all the sequencing centers especially the European Bioinformatics Institute and National Centre for Biotechnology Information. Thanks to the British Heart Foundation who funded the work in the laboratory of I.C.W. Thanks to Yves Dehouck at Université Libre Des Bruxelles for supplying us with the batch PoPMuSiC2.0 results of our variant datasets.

Funding: Biotechnology and Biological Science Research Council Research Development Fellowship (BB/D526502/1 to D.R.W.).

Conflict of Interest: none declared.

REFERENCES

Ackerman,M.J. *et al.* (2003) Ethnic differences in cardiac potassium channel variants: Implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome. *Mayo Clin. Proc.*, **78**, 1479–1487.

Adamczak,R. *et al.* (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* , **59**, 467–475.

Anantharam,A. *et al.* (2003) Pharmacogenetic considerations in diseases of cardiac ion channels. *J. Pharmacol. Exp. Ther.*, **307**, 831–838.

Biro,J.C. (2006) Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor. Biol. Med. Model.*, **3**, 15.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Brookes,A.J. (1999) The essence of SNPs. *Gene*, **234**, 177.
Burg,E.D. *et al.* (2010) Tetramerization domain mutations in KCNA5 affect channel kinetics and cause abnormal trafficking patterns. *Am. J. Physiol. Cell Physiol.*, **298**, C496–C509.
Calabrese,R. *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.
Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
Care,M.A. *et al.* (2007) Deleterious SNP prediction: be mindful of your training data! *Bioinformatics*, **23**, 664–672.
Clifford,R.J. *et al.* (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, **20**, 1006.
Crotti,L. *et al.* (2005) KCNH2-K897T is a genetic modifier of latent congenital long QT syndrome. *Circulation*, **112**, 1251–1258.
Dehouck,Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
Dukka Bahadur,K.C. and Livesay,D.R. (2008) Improving position-specific predictions of protein functional sites using phylogenetic motifs. *Bioinformatics*, **24**, 2308–2316.
Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
Editorial (2008) Prepare for the deluge. *Nat. Biotechnol.*, **26**, 1099–1099.
Gromiha,M.M. *et al.* (2010) Sequence and structural analysis of binding site residues in protein-protein complexes. *Int. J. Biol. Macromol.*, **46**, 187–192.
Hall,M. (2000) Correlation-based feature selection for discrete and numeric class machine learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 359–366.
Hall,M. *et al.* (2009) The WEKA Data Mining Software: an update. *SIGKDD Expl.*, **11**, 10–18.
Jackson,H. and Accili,E. (2008) Evolutionary analyses of KCNQ1 and HERG voltage-gated potassium channel sequences reveal location-specific susceptibility and augmented chemical severities of arrhythmogenic mutations. *BMC Evol. Biol.*, **8**, 188.
Jiang,R. *et al.* (2006) Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy. *BMC Bioinformatics*, **7**, 417.
Jiang,R. *et al.* (2007) Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *Am. J. Hum. Genet.*, **81**, 346–360.
John,G. and Langley,P. (1995) Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 338–345.
Kääb,S. and Schulze-Bahr,E. (2005) Susceptibility genes and modifiers for cardiac arrhythmias. *Cardiovas. Res.*, **67**, 397–413.
Kaneko,S. *et al.* (2002) Genetics of epilepsy: current status and perspectives. *Neurosci. Res.*, **44**, 11–30.
Kapa,S. *et al.* (2009) Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation*, **120**, 1752–1760.
Krishnan,V.G. and Westhead,D.R. (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.
Kubisch,C. *et al.* (1999) KCNQ4, a novel potassium channel expressed in sensory outer hair cells, is mutated in dominant deafness. *Cell*, **96**, 437.
Long,S.B. *et al.* (2007) Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. *Nature*, **450**, 376–382.
Makielski,J.C. (2006) SIDS: genetic and environmental influences may cause arrhythmia in this silent killer. *J. Clin. Investig.*, **116**, 297–299.
Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.
Ng,P. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
Pearl,J. (1984) *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
Poujois,A. *et al.* (2006) Chronic neuromyotonia as a phenotypic variation associated with a new mutation in the ;KCNA1; gene. *J. Neurol.*, **253**, 957–959.
Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
Ramensky,V. *et al.* (2002) Human non-synonymous SNPs: Server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
Remillard,C.V. *et al.* (2007) Function of Kv1.5 channels and genetic variations of KCNA5 in patients with idiopathic pulmonary arterial hypertension. *Am. J. Physiol. Cell Physiol.*, **292**, C1837–C1853.
Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
Scheffer,H. *et al.* (1998) Three novel KCNA1 mutations in episodic ataxia type I families. *Hum. Genet.*, **102**, 464–466.
Schwartz, P.J. *et al.* (2003) How really rare are rare diseases?: The intriguing case of independent compound mutations in the long QT syndrome. *J. Cardiovas. Electrophysiol.*, **14**, 1120.
Scicluna,B.P. *et al.* (2008) The primary arrhythmia syndromes: Same mutation, different manifestations. Are we starting to understand why? *J. Cardiovas. Electrophysiol.*, **19**, 445–452.
Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
Singh,A. *et al.* (2008) MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Res.*, **36**, D815–D819.
Splawski,I. *et al.* (2000) Spectrum of mutations in long-QT syndrome genes : KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. *Circulation*, **102**, 1178–1185.
Stead,L.F. *et al.* (2010) KvDB; mining and mapping sequence variants in voltage-gated potassium channels. *Hum. Mutat.*, **31**, 908–917.
Stenson,P.D. *et al.* (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.*, **1**, 13.
Sunyaev,S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
The ENCODE Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799.
The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789.
The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
Torkamani,A. and Schork,N.J. (2007) Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics*, **23**, 2918–2925.
Wedekind,H. *et al.* (2006) Sudden infant death syndrome and long QT syndrome: an epidemiological and genetic study. *Int. J. Legal Med.*, **120**, 129–137.
Yue,P. and Moult,J. (2006) Identification and analysis of deleterious human SNPs. *J. Mol. Biol.*, **356**, 1263.