BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 16 2014, pages 2324–2333
doi:10.1093/bioinformatics/btu297

Genetics and population analysis

Advance Access publication April 26, 2014

EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms

Trisevgeni Rapakoulia[1], Konstantinos Theofilatos[2], Dimitrios Kleftogiannis[1], Spiros Likothanasis[2], Athanasios Tsakalidis[2] and Seferina Mavroudi[2,3,*]

[1]King Abdullah University of Science and Technology (KAUST), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal 23955-6900, Saudi Arabia, [2]Computer Engineering and Informatics Department, University of Patras, Building B, Patras, 26504, Greece and [3]Department of Social Work, School of Health Sciences, Technological Institute of Western Greece, Patras, Greece

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Single nucleotide polymorphisms (SNPs) are considered the most frequently occurring DNA sequence variations. Several computational methods have been proposed for the classification of missense SNPs to neutral and disease associated. However, existing computational approaches fail to select relevant features by choosing them arbitrarily without sufficient documentation. Moreover, they are limited to the problem of missing values, imbalance between the learning datasets and most of them do not support their predictions with confidence scores.

Results: To overcome these limitations, a novel ensemble computational methodology is proposed. EnsembleGASVR facilitates a two-step algorithm, which in its first step applies a novel evolutionary embedded algorithm to locate close to optimal Support Vector Regression models. In its second step, these models are combined to extract a universal predictor, which is less prone to overfitting issues, systematizes the rebalancing of the learning sets and uses an internal approach for solving the missing values problem without loss of information. Confidence scores support all the predictions and the model becomes tunable by modifying the classification thresholds. An extensive study was performed for collecting the most relevant features for the problem of classifying SNPs, and a superset of 88 features was constructed. Experimental results show that the proposed framework outperforms well-known algorithms in terms of classification performance in the examined datasets. Finally, the proposed algorithmic framework was able to uncover the significant role of certain features such as the solvent accessibility feature, and the top-scored predictions were further validated by linking them with disease phenotypes.

Availability and implementation: Datasets and codes are freely available on the Web at http://prlab.ceid.upatras.gr/EnsembleGASVR/dataset-codes.zip. All the required information about the article is available through http://prlab.ceid.upatras.gr/EnsembleGASVR/site.html

Contact: mavroudi@ceid.upatras.gr

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

Received on October 14, 2013; revised on April 22, 2014; accepted on April 23, 2014

*To whom correspondence should be addressed.

1 INTRODUCTION

Understanding the relationship between genotype and phenotype is a fundamental problem in biology and biomedicine. Today, genome-wide sequencing combined with high-throughput platforms deliver significant improvements over older methods for identifying DNA sequence variations (Abecasis *et al.*, 2010). One of the most common types of genetic variation in humans is the single nucleotide polymorphisms (SNPs). A particular subcategory of SNPs called non-synonymous coding SNPs (nsSNPs) or missense SNPs refers to a single base substitution in a coding region that causes an amino acid substitution in the corresponding protein. The effects of nsSNPs in molecular function range from complete neutrality to disease susceptibility and lethality (Manolio *et al.*, 2009). Coding-region SNPs not only characterize human evolution and diversity (Goldstein and Cavalleri, 2005) but are also associated with drug sensitivity (Giacomini *et al.*, 2007) and disease susceptibility (Bell, 2004). Classifying nsSNPs according to their phenotypic effects has an important insinuation for understanding several diseases and exploring genetic ancestry, evolution and diversity among species (Cargill *et al.*, 1999).

Consequently, the effective characterization of polymorphic variations emerges as a challenging area of research. Testing experimentally the relationship between nsSNPs and diseases is not a trivial task and has several disadvantages in terms of cost and time (Valentini *et al.*, 2002). For this purpose, several computational methods have been developed (Thusberg *et al.*, 2011). From the algorithmic perspective, there are two types of existing methodologies. The first type concerns the sequence homology methods that apply conservation analysis of amino acids substitutions among evolutionarily related proteins (Ng and Henikoff, 2001; Thomas *et al.*, 2003). In principle, highly conserved residues in the polypeptide chain seem to be intolerant to amino acid substitutions, whereas positions with low degree of conservation allow more substitution without affecting protein functionality. The latter type regards Machine Learning (ML) methods that reformulate the characterization problem to a classification task and categorize polymorphisms into neutral or pathogenic based on structural, functional, sequential and evolutionary attributes. Up to now, several classification techniques have been applied including Artificial Neural Networks (Bromberg and Rost, 2007), Random Forests (Li *et al.*, 2009), Naïve Bayes classifiers

(Adzhubei *et al*., 2010) and Support Vector Machines (SVM; Acharya and Nagarajaram, 2012).

Although these methods have led to the characterization of a great number of SNPs and achieve high classification performance, they have so far been limited by several disadvantages (Thusberg *et al*., 2011). For instance, features used for classification are chosen arbitrarily without sufficient documentation (Hu and Yan, 2008). Furthermore, the feature selection process, which is of great importance for several bioinformatics problems (Saeys *et al*., 2007), is not systematized (Huang *et al*., 2010). Additionally, handling imbalanced pathogenic and neutral datasets is another obstacle that the existing methodologies fail to tackle efficiently (Wei and Dunbrack, 2013). Similar to other bioinformatics applications, the problem of missing values is also present and restricts the generalization ability of the developed models. The effective imputation of missing features without loss of information or expensive computations are aspects that require further consideration.

In this study, a novel ML methodology called EnsembleGASVR is introduced. EnsembleGASVR is an embedded classification technique, which combines an Adaptive Genetic Algorithm (GA) with nu-Support Vector Regression (nu-SVR), through an ensemble algorithmic framework. EnsembleGASVR predicts whether a given nsSNP is pathogenic or neutral, based on an extensive feature set that contains the most indicative features proposed in the literature. The feature selection component is integrated to the learning phase of the methodology and uses an adaptive GA for selecting relevant feature subsets. Simultaneously, the GA optimizes the nu-SVR parameters using a novel objective function tailored to the problem-related requirements. Also, to deal with the problem of missing values in the datasets, alleviate the class imbalance lying in the examined datasets and raise the algorithm's overall performance, EnsembleGASVR combines eight individual classification models that extend the proposed algorithm to function as an ensemble technique. Specifically, for the classification of a single SNP, only models for which this particular SNP has missing values below a predefined threshold are deployed to classify it. By training multiple classifiers and using a fitness function specialized for imbalanced bioinformatics datasets, we take advantage of our full dataset distribution, while at the same time we reduce the effects of class imbalance. The combined action of multiple classifiers enables the acquisition of an extremely robust regression technique, which could be used to classify SNPs and assign scores for every prediction. Moreover, by varying the ensemble classification threshold, EnsembleGASVR becomes tunable in terms of sensitivity and specificity.

Experimental results indicate that the proposed algorithmic technique outperformed existing methods in the datasets, which were used in this study, and leads to the automatic identification of a small and consistent subset of polymorphic features. EnsembleGASVR achieves a correct prediction rate of 87.45% and a geometric mean of 82%, tested on an independent human nsSNPs dataset. A set of four features appears to be present in all the eight extracted independent SVR models, and useful conclusions are made about their role in characterizing SNPs. Moreover, one newly introduced feature, Protein Essentiality, is selected in six of the eight SVR models indicating its importance in classifying nsSNPs.

2 METHODS

2.1 Datasets and features

The data generation process follows the methodology described in a comprehensive review of nine computational prediction methods (Thusberg *et al*., 2011). The pathogenic dataset was extracted from PhenCode database and contains non-synonymous SNPs that affect human phenotype (September 2012 registry) (Giardine *et al*., 2007). Only nsSNPs annotated as disease causing in SwissProt (Yip *et al*., 2008) were included. SwissProt provides high-quality curated information about potential relationship of missense mutations and diseases and enhances the pool of pathogenic samples in our collection. The neutral dataset contains nsSNPs that have not yet been associated with any disease according to dbSNP database (Sherry *et al*., 2001). dbSNP is freely accessible and provides data in a variety of formats. From the dbSNP FTP service, we downloaded the ASN.1 flat files (Build 135, September 2012 registry), containing detailed information about all human genetic variants, grouped according to their chromosomal site. Then, the entries were filtered to meet the following criteria: (i) only non-synonymous (missense) SNPs are contained; (ii) have been validated and have not been withdrawn (validated = YES, not-withdrawn); (iii) have not been annotated as pathogenic or probable-pathogenic; (iv) have minimum allele frequency >0.01; and (v) have been reported in at least 25 individuals (50 chromosomes). By applying these constraints, we tried to eliminate the number of disease-associated variants in our neutral dataset. dbSNP entries that contained links to Online Mendelian Inheritance in Man (OMIM) database were discarded. Finally, the two datasets were compared and duplicate records in both sets were removed, to minimize the probability of false-negative and false-positive cases in the neutral and pathogenic sets, respectively. We ended up with a pathogenic dataset of 17 743 nsSNPs, obtained from 2147 different protein sequences and a neutral dataset of 48 684 nsSNPs, obtained from 16 534 different protein sequences.

Regarding the feature set construction, we studied features proposed by nine computational methods (Thusberg *et al*., 2011). From this big pool of proposed characteristics, 87 representative attributes that have adequate documentation and describe the amino acid sequence, protein functionality, structural properties and evolutionary conservation were selected. The computation of the feature vector was accomplished deploying existing tools and scripts written by us for the cases where no publicly available program existed. Additionally, one new characteristic was also introduced: protein essentiality that describes whether the nsSNPs are obtained from essential proteins (DEG database data were used). The essential proteins are vital for the maintenance of life, as their malfunction can be fatal for the organism. It has been also indicated that mutations in the corresponding essential genes are directly related to the occurrence of many diseases (Furney *et al*., 2006). The protein domains feature, which refers to nsSNPs that belong to a conserved protein domain, was computed for the first time using the InterProScan v5 software package (Quevillon *et al*., 2005). The utilization of InterProScan decreased the feature's missing values compared with its calculation using only Pfam domains software (Punta *et al*, 2012). Protein domains are conserved protein regions that can perform their specific function independently or in coordination with neighboring domains. Therefore, the specific location of nsSNPs within the protein can be crucial for characterizing their impact (Zhang *et al*., 2009). Table 1 presents the feature set categories, and a detailed description of the features can be found in the Supplementary Table S1.

2.2 GASVR algorithm

EnsembleGASVR is an embedded classification system that uses a hybrid combination of GA and nu-SVR classifier. SVR classifiers are a common form of SVMs of broad applicability to many pattern recognition problems. In principle, SVR presents high classification performance and low complexity because it relies on a subset of training data and

Table 1. Sequential, structural, functional and evolutionary attributes considered in the present study (in bold the features introduced in the present work)

Category	Description	Number of features	Reference
Sequential information	Local sequence information	42	(Capriotti *et al*., 2006)
	Transition frequencies	12	(Bromberg and Rost, 2007)
Structural information	Secondary structure	3	(Petersen *et al*., 2009)
	Solvent accessibility	2	(Petersen *et al*., 2009)
	Transmembrane helices	1	(Krogh *et al*., 2001)
	Protein stability changes	1	(Cheng *et al*., 2006)
	Disordered regions	9	(Dosztányi *et al*., 2005)
Functional information	Signal peptide	1	(Petersen *et al*., 2011)
	DNA-binding sites	1	(Yan *et al*., 2006)
	Phosphorylation sites	6	(Blom *et al*., 1999)
	Protein domains	***1***	(Adzhubei *et al*,. 2010)
	Protein essentiality	***1***	***Newly introduced***
	Protein functional annotation based on Gene Ontology	1	(Calabrese *et al*., 2009)
Evolutionary information	PANTHER outputs	7	(Thomas *et al*., 2006)

uses a simplified cost function for building the model (Kwon and Moon, 2007). In contrast to regular SVM classifier, the nu-SVR outputs are real values that can be used as scores to support predictions. In addition, based on the application requirements, nu-SVR models become tunable in terms of sensitivity and sensitivity by modifying the decision threshold. GAs are stochastic meta-heuristic optimization algorithms that have been inspired by evolutionary biology principles (Holland, 1975). The most relevant aspect of GAs is their ability to explore efficiently large search spaces and identify possible solutions, without getting trapped in local optimal. Moreover, they present high abilities to explore the search space and locate near-to-optimal solutions. In EnsembleGASVR's first step, an adaptive GA is deployed to tune nu-SVR parameters, which are as follows: (i) classifier parameters *C and nu;* (ii) Radial basis kernel bandwidth *gamma; and* (iii) *classification threshold*. Additionally, the combination of nu-SVR and GA systematizes the feature selection process and identifies relevant subsets of features under the embedded feature selection setting.

In the GA framework, each chromosome is a binary string that encodes feature subset and parameters values. Specifically, a 142-bit string is used where 88 bits encode features, 10 bits represent each of the parameters nu and threshold (integer and decimal part), 14 bits correspond to gamma value and 20 bits are used for the parameter C. A rank-based roulette wheel selection method controls the selection of the best candidates in each GA generation (Hancock, 1994). This selection mechanism is preferred compared with the single roulette wheel selection to raise the selection pressure toward better solutions when all solutions of the population have present similar fitness values.

Elitism is used to force the best solution of each population to be selected at least once in the next generation. The evaluation of each chromosome in the population is performed according to the following fitness function:

$$Fitness = a \cdot Accuracy + b \cdot GeometricMean$$

$$-c \cdot 10^2 \cdot MSE - d \cdot \frac{1}{88} \cdot Features - e \cdot \frac{1}{4151} \cdot SupportVectors \quad (1)$$

where Accuracy is the nu-SVR's accuracy, GeometricMean refers to the geometric mean of sensitivity and specificity, MSE is the mean square of errors, Features represent the size of selected feature set and SupportVectors is the number of support vectors, included in the trained nu-SVR model. The geometric mean is used to handle the imbalance between positive and negative training examples (Akbani and Kwek, 2004). The term SupportVectors in the fitness function allows EnsembleGASVR to use technical information of the nu-SVR classification model, in contrast to other wrapper methods that have been recently proposed in the literature, which use only the classifier's performance. Specifically, smaller and simpler SVR models are preferred, as they present better generalization properties. The range of the examined variables in the proposed fitness function where Accuracy $\in$ [0, 1], GeometricMean $\in$ [0, 1], MSE $\in$ [0, 0.01], Features $\in$ [1-Max_Features], where Max_Features is the maximum number of features that can be selected by our method (in our case 88), and SupportVectors $\in$[1, Training_Size], where Training_Size is the number of training examples. These variables were multiplied with certain constants to normalize their values to range from 0 to 1. The constants a, b, c, d and e in Equation (1) are user-specified weights. The following values were assigned without experimentation: a = 0.5, b = 0.5, c = 0.01, d = 0.005 and e = 0.001. These values were selected to reflect the priorities of the goals for classifying missense SNPs. Using these values, EnsembleGASVR achieves high classification performance and simultaneously generates a simple and effective model. The order of our objectives significance may be extracted by observing the weight value of each separate goal. So, the classification accuracy, the geometric mean and the MSE of the classifier are the most significant, with the number of selected features being less significant and the number of the support vectors being the least significant objective. To avoid overfitting problems, we did not attempt to optimize these values.

Then the differentiation operators, crossover and mutation are applied to the top-ranked candidate solutions to create a new population. The crossover operator applies 2-point crossover to obtain a new offspring from two parents. In our current implementation, the crossover rate is constant and set to 0.9. This rate increases the variability of the crossover operator and with a small probability candidate solutions pass to the next generation unaffected. This property is essential when good solutions emerge in early stages of the algorithm.

To achieve the desired trade-off between exploration and exploitation, EnsembleGASVR uses an adaptive mutation rate. The mutation rate is

determined by using the following equation:

$$P_m(n) = \begin{cases} 0.2 - n \cdot \dfrac{0.2 - \dfrac{1}{P_S}}{MAX_G}, A\_S < 90\% \\ 0.2 + n \cdot \dfrac{0.2 - \dfrac{1}{P_S}}{MAX_G}, A\_S >= 90\% \end{cases} \quad (2)$$

where n is the current generation, P_S is the population size and MAX_G is the maximum generation, as defined by the termination criteria, and A_S is the average similarity between the members of the population and its best member. The proposed adaptive mutation operator attempts to perform global optimization in its first iterations while gradually switching its behavior to local optimization. The value 0.2 was selected in the Equation (2) as an extremely high-mutation probability value that transforms the proposed algorithm to a random search algorithm in its first steps. To avoid getting trapped in local optima, the average similarity of each chromosome with the best individual of the population is estimated in each generation. If this average similarity exceeds 90%, the mutation probability is increased by a factor of $(0.2–1/P_S)/(MAX_G)$, instead of being reduced. When the average similarity of the chromosomes within a population is over 90%, there is a high possibility that the proposed algorithm has stacked in a local optimal solution.

The overall flowchart of the GASVR algorithm is presented in Figure 1A.

The size of the initial population was set to 30 chromosomes. Further experiments using the training set were conducted to optimize this parameter, but the results were not improved significantly. The termination criterion was the convergence of the population. The population is considered converged when the average fitness across the current population is <5% away from the best fitness of the current population. The maximum number of generations was set to 150 after observing that most of the proposed algorithms executions were terminated in ~100–120 generations when the convergence criterion was activated.

EnsembleGASVR is implemented in Matlab version R2011a, and all source code and datasets are available at http://prlab.ceid.upatras.gr/EnsembleGASVR/dataset-codes.zip.

2.3 EnsembleGASVR

To alleviate the problem of missing values, to take advantage of our full imbalanced dataset and to improve the overall classification performance, the aforementioned algorithm was extended to function as an ensemble technique, named EnsembleGASVR. Ensemble methodologies (Rokach, 2009) are used to increase the effectiveness of individual classifiers and maximize the performance by training multiple classifiers and combining their decisions into a single output. The main profit of ensemble methodologies is the improvement of efficiency and accuracy: each classifier introduces errors, but the combination of different classifiers, as they have been trained on different subsets of the original data, outperforms a single classifier (Kittler *et al.*, 1998). The development of several classifiers for each sub-training set is not affected by the heterogeneity of the whole imbalanced training set in contrast to a single classifier trained with whole training examples (Chang, 2003). The problem of missing values is

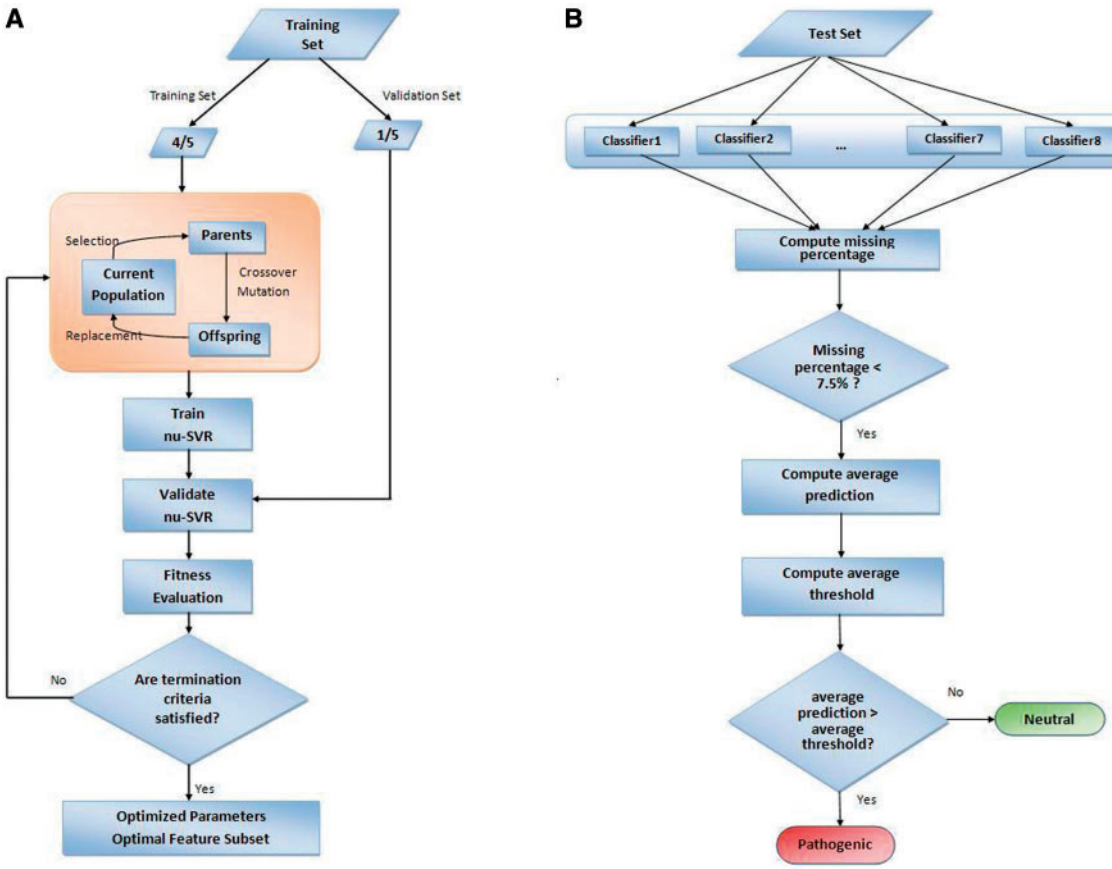

Fig. 1. (**A**) GASVR algorithm flowchart: an embedded feature selection approach that uses a hybrid combination of GA and nu-SVR classifier. (**B**) EnsembleGASVR flowchart: nu-SVR models satisfying missing values criteria are combined to produce a single global prediction

also tackled effectively owing to the extraction of classifier-specific features subsets. The final prediction of test samples takes into account only those classifiers that do not have high percentage of missing variables in their feature vector. In contrast to existing methodologies, which discard data samples or use complicated imputation techniques, the aforementioned approach provides significant computational savings, and offers an elegant way of imputing missing values, without any loss of information.

Taking advantage of the initial dataset for constructing more efficient classifiers, we divided randomly the initial dataset into two equal size sets for training and testing, both containing 33 208 nsSNPs. The training set was further partitioned randomly into n subsets of equal size. The random partitioning procedure ensures that the ratio between positive and negative samples was kept the same as in the original dataset. A classifier was trained for each training subset, applying the hybrid training process described in the previous subsection. The number of independent classifiers (n) was selected through an internal trial and error procedure that attempted to use most of the available features (>95% of them) in at least one classifier, while maintaining a reasonable sample size (>1000 samples for the minority class). In specific, the number of the classifiers, starting with two independent classifiers, was in each step increased by one until the above criteria are satisfied. This fact enabled the proposed methodology to use most of the available information, extracting a set of diversified classifiers. The minimum number of independent classifiers (n) that satisfied the above criteria has been found equal to eight. The size of the training subset for this setting (1109 pathogenic and 3042 neutral examples, respectively), which was used to train every independent classifier, was sufficient to avoid the problem of overfitting because of using small datasets (Lenth, 2001), (Schaaf *et al.*, 2012). The eight nu-SVR models were combined to form an ensemble classification model, which can be applied to predict the pathogenicity of new and unseen nsSNPs. Regarding the testing process, the new instances are classified by the eight nu-SVR models, and the ensemble majority voting process is applied. A prediction value, a classification threshold and the feature subset used for classification are returned after the execution of each of the eight classifiers. However, if a feature subset of a classifier contains features with missing values, in a percentage >7.5%, then the output of the classifier is considered as low confident and is not taken into account for the final prediction (Fig. 1B). The threshold for discarding a classification model because of missing values was optimized experimentally using only the training dataset. In general, the proposed methodology uses models with a small percentage of missing values. These missing values are calculated using the K-Nearest Neighbor (KNN)-impute methodology (Troyanskaya *et al.*, 2001). Finally, an average prediction value and an average classification threshold are calculated based on the output of those classifiers that satisfy the above criterion. If the average prediction value is higher than the classification threshold, EnsembleGASVR predicts nsSNP as pathogenic, otherwise it predicts it as neutral. The way the eight individual models are combined to classify a new nsSNP is summarized in Table 2.

Table 2. EnsembleGASVR flowchart (in bold the features introduced in the present work)

EnsembleGASVR Algorithm

```
for each test example do:
    for i:←1 to 8 do:
      compute prediction[i]
      compute threshold[i]
      compute missing_percentage[i]

    if missing_percentage[i] < 7.5% then
        total_prediction ←total_prediction + prediction[i]
        total_threshold ←total_threshold + threshold[i]
        total_classifiers ←total_classifiers + 1
      end if
    end for
    average_prediction ← total_prediction / total_classifiers
    average_threshold ←total_threshold / total_classifiers

    if average_prediction > average_threshold then
      Classify test example as pathogenic
    else
      Classify test example as neutral
    end if
end for
```

3 RESULTS AND DISCUSSION

3.1 Comparison with existing methods

The quality of EnsembleGASVR prediction was evaluated according to seven metrics: accuracy, sensitivity, specificity, geometric mean of sensitivity and specificity, Matthews Correlation Coefficient (MCC), precision and area under the curve (AUC). It is evident that the values of sensitivity and specificity of a classifier depend on the specified classification threshold between neutral and pathogenic classes. In our case, the regression nature of nu-SVR models enables EnsembleGASVR model to be flexible in terms of sensitivity and specificity. By changing the classification threshold, different rates between sensitivity and specificity can be achieved. For the case of imbalance learning, the geometric mean of sensitivity and specificity is the most adequate metric to assess the performance of a classifier (Kubat and Matwin, 1997).

EnsembleGASVR was compared with the following four well-known predictors of pathogenic nsSNPs: MutationAssessor, Polyphen-2, PANTHER and SIFT. The main reason we chose the aforementioned methods for our benchmarking, is that they are widely used, frequently updated and also enable real-time analysis of large lists of nsSNPs (Castellana and Mazza, 2013). MutationAssessor computes a functional impact score and estimates the effects of mutations using evolutionary conservation criteria (Reva *et al.*, 2011). This score derives from a set of evolutionarily conserved residues that are computed by clustering into subfamilies multiple alignments of homologous sequences. Polyphen-2 (Adzhubei *et al.*, 2010) uses a Naïve Bayes classifier that has been trained using two different datasets (HumVar and HumDiv) and it is up to users which of the two training sets will be selected for the classification. In this study, Polyphen-2's performance was evaluated, deploying both its alternative training sets. SIFT uses a homology-based approach to predict whether an amino acid substitution affects protein functionality and alters the phenotype (Ng and Henikoff, 2001, 2002). SIFT

aligns a query protein with homologous sequences, using the PSI-BLAST algorithm, and then it assigns a score in every residue, according to the probability of the occurrence of the given residue at the specific position in the alignment. Mutations in highly conserved positions more likely lead to disease-associated phenotypes, whereas mutations in not-conserved sites have neutral effects. Similarly, PANTHER (Thomas *et al.*, 2003) predicts the impact of nsSNPs based on position-specific evolutionary conservation scores, obtained from multiple sequence alignments of related proteins. As evolutionary information was incorporated in our classification, using features derived from the PANTHER output, PANTHER method provides a baseline measure for our benchmarking.

To present a fair comparison with respect to all the other methods, we tested the performance on two independent datasets: (i) HumVar: containing 22 196 deleterious and 21 119 neutral mutations in 9679 human proteins. (ii) Independent test set derived from our data collection: because our data collection is large enough, we chose the simple hold-out approach to partition the data randomly to completely independent training and testing sets. Our final testing set contains 33 213 nsSNPs from 18 681 human proteins. The two different test sets share 16 798 common nsSNPs, while the number of overlapping proteins is 6538. More information about the overlap of the used independent test set with the HumVar dataset are provided in the Supplementary Table S5.

To assess the robustness of the proposed algorithm approach, EnsembleGASVR was run 10 times, training 10 different models. During the training procedure, the training set was partitioned randomly into eight equal-size subsets. The average classification performance and the standard deviation of the trained models are presented in Tables 3 and 4.

From the comparison, we excluded HumDiv dataset that contains non-human mammalian homologs as negative sample. The non-human HumDiv negative dataset has different characteristics compared with the human ones, and the training and testing of our method in such a dataset is out of the present article's scope. However, Polyphen-2's latest version presented significant advantages and reported high classification performance on that dataset.

Figures 2 and 3 present the receiver-operating characteristic (ROC) curves analysis for both studied datasets. Extended experimental results are included in Supplementary Tables S3 and S4 containing classification performance of all the studied methods in HumDiv as well as the accuracy for the latest dbSNP registry (build 139). Apparently, EnsembleGASVR outperforms the other methods in accuracy, specificity, geometric mean, precision, MCC and AUC in both datasets. Polyphen-2, on the other hand, achieves higher sensitivity in both datasets but much lower performance in all other examined metrics. That fact indicates that EnsembleGASVR tackles more effectively the class imbalance problem. As it has already been mentioned, the geometric mean gives better and more reliable insight on the predictive power of a classifier. One reason for this advantage of our method is its ability to combine features disregarding of the existence of mutual information among them. In opposite, Polyphen is based on the Naive Bayes classifier, which assumes independence between the deployed features. Moreover, we observe that PANTHER and SIFT that rely on alignment with homolog proteins achieve high specificity in both datasets but much lower sensitivity. That fact can be attributed to the absence of homologs in their libraries for all of the tested cases.

To verify the informative power of specific attributes, the feature subsets used by each individual classifier were compared and those that presented great discriminatory ability were identified. Relative Solvent Accessibility, probability of wild-type residue in PANTHER library (P_{wt}), message column of PANTHER output and protein functional annotation, based on Gene Ontology terms (Gene Ontology log-odds (LGO) score LGO), were used by all the classifiers, and therefore, they present great discriminatory ability. It is also noteworthy that the attribute (protein essentiality) was included in six of the eight extracted feature subsets. This observation denotes the discriminative ability of this feature, highlights its effectiveness and recommends its usage in the nsSNPs classification problem. More details about the selected features of each individual classifier can be found in Supplementary Table S2.

3.2 Exploring the value of the predicted score

It was mentioned previously that SVRs, when used for classification problems, support prediction with real value scores that could be exploited as confidence scores (extreme high and low values indicate more reliable predictions and values near zero represent uncertain predictions of low confidence). To validate this finding, we obtained the confidence scores of the EnsembleGASVR model that achieved the highest geometric mean value in the training set, and we studied the top-10–scored predictions coming from our independent test set. We found that all these predictions belong to the class of diseases-associated nsSNPs, and all of them are reported in the dbSNP database. The top-10–scored mutations are presented in Table 5. From this table, it is observed that all these mutations have been connected with specific pathogenesis. A more detailed search in OMIM database revealed that all of the top 10 predictions are associated with rare single gene diseases and six of them belong to the class of autosomal disorders. Moreover, 4 of the 10 mutated proteins are expressed highly in vital organs such as liver, brain and colon. Regarding their functional effects, 50% is lethal and death occurs in the early childhood. The rest cause severe syndromes, such as Li-Fraumeni syndrome and Xeroderma pigmentosum, and in some cases the risk of developing cancer is highly increased. These observations demonstrate that EnsembleGASVR prediction scores are extremely effective in measuring the confidence of the derived predictions.

Then, as an alternative validation procedure, we studied whether these predictions are associated with genome-wide association studies (GWAS) based on the latest catalog (https://www.genome.gov/26525384). As the scope of this article is to study nsSNPs, the GWAS SNPs were filtered to keep only the non-synonymous ones excluding the intergenic and intronic SNPs. GWAS latest version (January 2014) includes 628 nsSNPs. The whole batch of our predictions was scanned, and 198 matches were found achieving a statistically significant coverage of 31.52%.

Next, we investigated the possibility that the extracted confidence score could act as a metric for the severity of the polymorphism for the disease it causes. To achieve this, we used the

Table 3. Comparison of the studied methods on EnsembleGASVR indpendent test set (HumVar/HumDiv labels refer to the specific classification models used). (in bold the features introduced in the present work)

Method	Accuracy	Sensitivity	Specificity	Geometric mean	Precision	MCC	AUC
Mutation Assessor	0.838	0.728	0.878	0.799	0.685	0.596	0.849
SIFT	0.693	0.378	0.808	0.553	0.418	0.193	0.834
PANTHER	0.744	0.312	0.901	0.531	0.535	0.261	0.572
Polyphen-2 (HumVar)	0.791	0.807	0.785	0.796	0.578	0.542	0.872
Polyphen-2 (HumDiv)	0.731	**0.850**	0.688	0.765	0.498	0.478	0.853
EnsembleGASVR	**0.875 (±0.0004)**	0.713 **(±0.0051)**	**0.934 (±0.0025)**	**0.816 (±0.0018)**	**0.798 (±0.0054)**	**0.672 (±0.0009)**	**0.906 (±0.0063)**

Table 4. Comparison of the studied methods on HumVar dataset (HumVar/HumDiv labels refer to the specific classification models used). (in bold the features introduced in the present work)

Method	Accuracy	Sensitivity	Specificity	Geometric mean	Precision	MCC	AUC
Mutation Assessor	0.804	0.768	0.842	0.804	0.836	0.611	0.878
SIFT	0.626	0.368	**0.896**	0.574	0.788	0.309	0.861
PANTHER	0.592	0.309	0.889	0.524	0.745	0.242	0.522
Polyphen-2 (HumVar)	0.818	0.825	0.811	0.818	0.821	0.636	0.893
Polyphen-2 (HumDiv)	0.798	**0.872**	0.720	0.792	0.766	0.600	0.877
EnsembleGASVR	**0.818 (±0.002)**	0.762 **(±0.0133)**	0.877 **(±0.009)**	**0.817 (±0.0028)**	**0.866 (±0.0063)**	**0.642 (±0.0029)**	**0.902 (±0.002)**

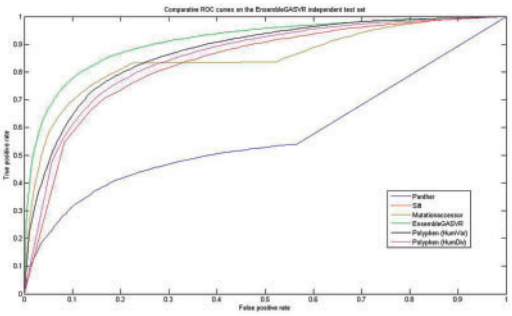

Fig. 2. Comparative ROC curves for the produced independent test set

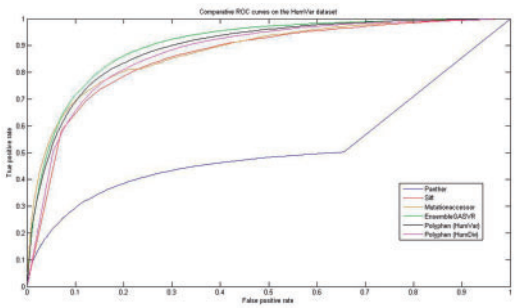

Fig. 3. Comparative ROC curves for HumVar dataset

annotations provided by UniProt for the pathogenic SNPs. In particular, the pathogenic SNPs can be characterized as Severe, Moderately Severe, Intermediate form, moderate, Mild and Not Characterized. We intended to compare the 1000 highest scored pathogenic SNPs with the 1000 lowest scored pathogenic diseases. However, only 49 among the highest scored and 58 SNPs among the lowest scored were actually characterized for their severity. Because of the limited data samples, an additional binning of characterization categories was required. The Severe, Moderately Severe and Intermediate characterization categories were combined to form the first group, and the Mild and Moderate characterizations formed the second one. The results are presented in Table 6. Applying the Fisher's Exact statistical test, a statistical significant difference ($P = 0.01611$) was uncovered between the two groups of mutations. Despite this significant finding, it is obvious that the limited total number of the severity characterized SNPs does not allow for more general conclusions and more firm findings could be extracted as this number rises.

4 CONCLUSIONS AND FUTURE WORK

A novel computational framework was introduced for predicting neutral and pathogenic polymorphic variations. The important component of the proposed methodology is the utilization of a two-phase algorithm that combines multiple nu-SVR classifiers under the ensemble setting. The utilization of a GA optimizes the ensemble framework parameters, the nu-SVR classifier parameters and reveals compact feature subsets. Moreover, a problem-specific scoring function was introduced that maximizes the classification performance and simultaneously produces interpretable and relative simple classification models. Although, all

Table 5. The top-10–scored nsSNPs, according to EnsembleGASVR confidence score and the associated diseases

UniProt ID	Sequence position	Tissue specificity	Disease	Reference
P00740	435	Synthesized in the liver and secreted in plasma	Hemophilia B(HEMB)	(Espinós *et al*., 2003)
P53634	301	Ubiquitous	Papillon-Lefevre syndrome (PLS)	(Hart *et al*., 2000)
P21817	4793	Skeletal muscle and brain	Central core disease of muscle (CCD)	(Monnier, 2001)
P45381	21, 24	Brain white matter	Canavan disease (CAND)	(Sistermans *et al*., 2000)
P18074	673	None reported	Trichothiodystrophy photosensitive (TTDP)	(Botta *et al*., 1998)
Q9Y6Q6	175	Ubiquitous	Osteopetrosis, autosomal recessive 7 (OPTB7)	(Guerrini *et al*., 2008)
P16144	325	Expressed by epithelia colon and placenta epidermis lung duodenum	Epidermolysis bullosa letalis, with pyloric atresia (EB-PA)	(Nakano *et al*., 2001)
P28715	858	Non reported	Xeroderma pigmentosum complementation group G (XP-G)	(Lalle *et al*., 2002)
P15848	146	None reported	Mucopolysaccharidosis 6 (MPS6)	(Simonaro and Schuchman, 1995)
P04637	196	Ubiquitous	Li-Fraumeni syndrome (LFS)	

Table 6. Comparing the severity of top 1000 scored and bottom 1000 scored pathogenic SNPs

Mutation set	Severe, moderately severe intermediate form	Mild, moderate	Not characterized
1000 Highest scored pathogenic SNPs	27	22	951
1000 lowest scored Pathogenic SNPs	20	38	942

these concepts are not new in the field of pattern recognition, this is the first time they are deployed in one unified framework that comfortably predicts nsSNPs and outperforms well-known existing methodologies. The performance superiority can be attributed to the ensemble nature of the method, which achieves high classification performance, rebalances the learning datasets, ignores overfitting issues and presents great generalization capabilities. In addition, the combination of decisions coming from different classifiers provides an elegant way of manipulating missing values without loss of information. Except for the problem of classifying missense SNPs, the proposed computational technique is general enough and it can be applied to different bioinformatics problem where the class imbalance and the missing values problems are present. Besides, an extensive study was performed for the most relevant features to the problem of classifying SNPs, and an 88-features superset was proposed. This superset contains attributes, coming from several categories, which capture all the adequate information of functional proteins and can characterize efficiently pathogenic or neutral variations. In contrast to the *ad hoc* selection of features, the systemization of the feature selection process is a valuable tool for linking specific protein characteristics with functionality and pathogenicity. The embedded feature selection component of EnsembleGASVR revealed a small feature subset, which is stable, and always present in the ensemble models with 100% frequency. This subset contains four features mostly related to evolutionary and functional information. Furthermore, the new feature that describes vital properties was selected with 75% frequency. Regarding the non-selected features, only two features (whether the nsSNP occurs in the close environment of T-phosphorylation site and whether the wild-type or mutant residue is Valine) were not selected, and this could be attributed to the mutual information shared by many of the examined features. Another significant contribution of this work is the publication of a high-quality novel dataset. Moreover, the implementation of the proposed methodology is publicly available and can be run on commodity computers in a reasonable amount of time. Nonetheless, there still is room for improvement. The main drawback of EnsembleGASVR is the slow execution time required for training (up to 24 h for an i5 processor workstation). However, this is not a restricting factor for the application of the trained model, as it requires only 15–25 s to classify a single new SNP. The high-computational cost of the training phase is attributed to the current implementation of the ensemble framework and the GA implementation, which is sequential. However, the ensemble nature of EnsembleGASVR and the evolutionary properties of the GA optimizer are ideal for parallelization. In our ongoing research, we are developing a master/slave program that can run in multicore architectures and can take advantage of cloud-oriented infrastructures. Another limiting factor is the lack of well-characterized non-damaging human homologs datasets that can serve as negative examples. Another interesting future idea is the prediction of disease-related intronic and intergenic SNPs using GWAS data. Regarding the biological part of this work, our future plan is to study more extensively the linkage between the confidence scores of our predictor and the pathogenicity of SNPs. The preliminary results for this task indicated that our predicted

confidence score is related to the severity of pathogenic SNPs. As new data are made available through studies such as the GWAS, this finding could be reinforced and new specific methods and tools could be implemented for this purpose.

ACKNOWLEDGEMENTS

The authors thank the Pattern Recognition Laboratory (http://prlab.ceid.upatras.gr) of the University of Patras, Greece, for hosting our codes and datasets.

Funding: Trisevgeni Rapakoulia and Dimitrios Kleftogiannis were supported by the King Abdullah University of Science and Technology (KAUST).

Conflict of Interest: none declared.

REFERENCES

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

Acharya,V. and Nagarajaram,H.A. (2012) Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Hum. Mut.*, **33**, 332–337.

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

Akbani,R. and Kwek,S. (2004) Applying support vector machines to imbalanced datasets. *Lect. Notes Comput. Sci.*, **3201**, 39–50.

Bell,J. (2004) Predicting disease using genomics. *Nature*, **429**, 453–456.

Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Botta,E. *et al.* (1998) Analysis of mutations in the XPD gene in Italian patients with trichothiodystrophy: site of mutation correlates with repair deficiency, but gene dosage appears to determine clinical severity. *Am. J. Hum. Genet.*, **63**, 1036–1048.

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

Calabrese,R. *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.

Capriotti,E. *et al.* (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.

Cargill,M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.

Castellana,S. and Mazza,T. (2013) Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief. Bioinform.*, **4**, 448–459.

Chang,Y.I. (2003) *Boosting SVM Classifiers with Logistic Regression*. Technical report, Academia Sinica, 2003 [http://www3stat.sinica.edu.tw/library/c_tec_rep/2003-03.pdf].

Cheng,J. *et al.* (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.

Dosztányi,Z. *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Espinós,C. *et al.* (2003) Molecular analyses in hemophilia B families: identification of six new mutations in the factor IX gene. *Haematologica*, **88**, 235–236.

Furney,S.J. *et al.* (2006) Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics*, **7**, 165.

Giacomini,K.M. *et al.* (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin. Pharmacol. Ther.*, **81**, 328–345.

Giardine,B. *et al.* (2007) PhenCode: connecting ENCODE data with mutations and phenotype. *Hum. Mut.*, **28**, 554–562.

Goldstein,D.B. and Cavalleri,G.L. (2005) Genomics: understanding human diversity. *Nature*, **437**, 1241–1242.

Guerrini,M.M. *et al.* (2008) Human osteoclast-poor osteopetrosis with hypogammaglobulinemia due to TNFRSF11A (RANK) mutations. *Am. J. Hum. Genet.*, **83**, 64–76.

Hart,P.S. *et al.* (2000) Identification of cathepsin C mutations in ethnically diverse papillon-Lefèvre syndrome patients. *J. Med. Genet.*, **37**, 927–932.

Holland,J.H. (1975) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, Oxford, England.

Hu,J. and Yan,C. (2008) Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information. *BMC Bioinformatics*, **9**, 297.

Huang,T. *et al.* (2010) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS One*, **5**, e11900.

Kittler,J. *et al.* (1998) On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, 226–239.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Kubat,M. and Matwin,S. (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: *Proceedings of the 14th International Conference on Machine Learning*. pp. 179–186.

Kwon,Y.K. and Moon,B.R. (2007) A hybrid neurogenetic approach for stock forecasting. *IEEE Trans. Neural Netw.*, **18**, 851–864.

Lenth,R. (2001) Some proctical guidelines for effective sample size determination. *Am. Stat. J.*, **55**, 187–193.

Lalle,P. *et al.* (2002) The founding members of xeroderma pigmentosum group G produce XPG protein with severely impaired endonuclease activity. *J. Invest. Dermatol.*, **118**, 344–351.

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

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Monnier,N. (2001) Familial and sporadic forms of central core disease are associated with mutations in the C-terminal domain of the skeletal muscle ryanodine receptor. *Hum. Mol. Genet.*, **10**, 2581–2592.

Nakano,A. *et al.* (2001) Epidermolysis bullosa with congenital pyloric atresia: novel mutations in the beta 4 integrin gene (ITGB4) and genotype/phenotype correlations. *Pediatric Res.*, **49**, 618–626.

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

Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446. doi:10.1101/gr.212802.

Petersen,B. *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Quevillon,E. *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118. doi:10.1093/nar/gkr407.

Rokach,L. (2009) Ensemble-based classifiers. *Artif. Intell. Rev.*, **33**, 1–39.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Schaaf,A. *et al.* (2012) Multivariate modeling of complications with data driven variable selection: guarding against overfitting and effects of data set size. *Radiother. Oncol.*, **105**, 115–121.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Simonaro,C.M. and Schuchman,E.H. (1995) N-acetylgalactosamine-4-sulfatase: identification of four new mutations within the conserved sulfatase region causing mucopolysaccharidosis type VI. *Biochim. Biophys. Acta*, **1272**, 129–132.

Sistermans,E.A. *et al.* (2000) Mutation detection in the aspartoacylase gene in 17 patients with Canavan disease: four new mutations in the non-Jewish population. *Eur. J. Hum. Genet.*, **8**, 557–560.

Thomas,P.D. *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

Thomas,P.D. *et al.* (2006) Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.*, **34**, W645–W650.

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

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Valentini,G. *et al.* (2002) Structure and function of human erythrocyte pyruvate kinase. Molecular basis of nonspherocytic hemolytic anemia. *J. Biol. Chem.*, **277**, 23807–23814.

Wei,Q. and Dunbrack,R.L. (2013) The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS One*, **8**, e67863.

Yan,C. *et al.* (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 262.

Yip,Y.L. *et al.* (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mut.*, **29**, 361–366.

Zhang,R. and Lin,Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.*, **37**, D455–D458.