Received: 7 January 2019 | Revised: 1 April 2019 | Accepted: 3 May 2019

DOI: 10.1002/humu.23785

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

SPECIAL ARTICLE

Predicting the change of exon splicing caused by genetic variant using support vector regression

Ken Chen[1] | Yutong Lu[1] | Huiying Zhao[2] | Yuedong Yang[1,3]

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China

[3]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, Guangzhou, China

Correspondence
Yuedong Yang, School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China.
Email: yangyd25@mail.sysu.edu.cn
Huiying Zhao, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China. Email: zhaohy8@mail.sysu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61772566, 81801132, U1611261; Guangdong Introducing Innovative and Entrepreneurial Teams, Grant/Award Number: 2016ZT06D211; Guangdong Province Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Grant/Award Number: 2017B030314026; NIH, Grant/Award Number: U41 HG007346; CAGI conference, Grant/Award Number: NIH R13 HG006650

Abstract

Alternative splicing can be disrupted by genetic variants that are related to diseases like cancers. Discovering the influence of genetic variations on the alternative splicing will improve the understanding of the pathogenesis of variants. Here, we developed a new approach, PredPSI-SVR to predict the impact of variants on exon skipping events by using the support vector regression. From the sequence of a particular exon and its flanking regions, 42 comprehensive features related to splicing events were extracted. By using a greedy feature selection algorithm, we found eight features contributing most to the prediction. The trained model achieved a Pearson correlation coefficient (PCC) of 0.570 in the 10-fold cross-validation based on the training data set provided by the "vex-seq" challenge of the 5th Critical Assessment of Genome Interpretation. In the blind test also held by the challenge, our prediction ranked the 2nd with a PCC of 0.566 that demonstrates the robustness of our method. A further test indicated that the PredPSI-SVR is helpful in prioritizing deleterious synonymous mutations.

The method is available on https://github.com/chenkenbio/PredPSI-SVR.

KEYWORDS

alternative splicing, splice site motif, support vector machine, synonymous mutation, Vex-seq

1 | INTRODUCTION

Human genes often express pre-mRNAs containing multiple introns and exons. Usually, introns will be spliced out with exons that are connected to form mature mRNAs. The alternative splicing of exons allows pre-mRNA to be spliced into diverse mature mRNAs (G.-S. Wang & Cooper, 2007). Thus, alternative splicing greatly contributes to the complexity of the human genome, and allows to generate protein isoforms with different functions expressed from one gene (Baralle & Giudice, 2017). The changes of alternative splicing have been widely known to relate to human diseases, and even cancers (Climente-González, Porta-Pardo, Godzik, & Eyras, 2017). For example, variants occurring around splice sites can cause Birt-Hogg-Dubé syndrome, cystic fibrosis, Duchenne muscular dystrophy, and others (Anna & Monika, 2018; Furuya et al., 2018). Importantly, many synonymous mutations happening in exons that do not change encoded proteins were found to influence gene functions (Goodman, Church, & Kosuri, 2013; Parmley, Chamary, & Hurst, 2006), or act as driver mutations in cancers due to their associations with splicing changes (Supek, Miñana, Valcárcel, Gabaldón, & Lehner, 2014).

Genetic variants that affect splicing events usually alter splicing signals in pre-mRNAs. The most fundamental splicing signals are located in 5′ splice sites, 3′ splice sites, and branch point sequences (Will & Lührmann, 2011). Usually, 5′ splice sites start with "GU" and 3′ splice sites end with "AG," marking the beginning and end of introns, respectively. On the other hand, branch point sequences locate near the upstream of 3′splice site in introns, which helps to form lariat-like intermediates for introns that are spliced out. In addition, splicing regulatory elements are also required to precisely identify splice sites existing in exons and introns, including exonic splicing enhancer and silencer (ESE/ESS), intronic splicing enhancer and silencer. These

regulatory elements are short sequences in pre-mRNAs that can modulate alternative splicing by interacting with regulatory proteins (Wang & Burge, 2008). Apart from splicing signals in pre-mRNA sequences, the secondary structure of pre-mRNAs can affect splicing as well (McManus & Graveley, 2011).

A common form of alternative splicing in mammals is exon skipping, where an exon will be spliced into mature mRNA or skipped entirely (Katz, Wang, Airoldi, & Burge, 2010). The skipping event of an exon is often measured by the percentage of the exon to be spliced in, namely PSI or Ψ, and the difference of Ψ ($\Delta\Psi$) can be used to quantify the change of exon splicing. To quantify the alternative splicing, Xiong et al. (2015) used a high-throughput sequencing technique to measure genome-wide exon splicing, from which they have designed a method SPANR to predict Ψ based on a deep Bayes network. Though the method was able to obtain $\Delta\Psi$ by predicting Ψ values individually for wild-type (WT) sequences and their genetic variants, the indirect way to predict $\Delta\Psi$ is usually less accurate compared to methods specifically designed for the prediction. At the same time, Rosenberg, Patwardhan, Shendure, & Seelig (2015) designed a new method HAL to predict $\Delta\Psi$ by using hexamer motifs of splicing patterns trained from more than two million synthetic mini-genes. However, the method can only make predictions for variants occurring in exons or splice donors (introns within 6 bp from the 5′ splice sites), but not in other regions. In addition, this method does not consider other affecting factors.

Recently, Adamson, Zhan, & Graveley (2018) used a novel experimental technique, variant exon sequencing (vex-seq), to measure the impact of genomic variants on alternative splicing that are hard to be detected by traditional approaches using poly(A)+RNA-seq alone. Vex-seq adopts a barcoding approach and is able to detect variants in exons and flanking introns. This method was applied on 2,059 variants, and has produced a precise data set for $\Delta\Psi$ caused by each variant. On the data set, the method SPANR achieved a low correlation while the method HAL could not make predictions for mutations outside exons and donor regions. Thus, this Vex-seq data set is valuable for developing an accurate model for predicting $\Delta\Psi$.

Here, we present a new method (viz., PredPSI-SVR) that uses support vector regression for predicting $\Delta\Psi$ caused by variants. This method was trained on selected features, including DNA sequence, DNA conservation score, splicing site, splicing regulatory elements, and mRNA secondary structure. The 10-fold cross-validation test indicated that the method outperformed SPANR. It was ranked the 2nd with a Pearson correlation coefficient (PCC) of 0.566 on the blind prediction for vex-seq competition among the 5th Critical Assessment of Genome Interpretation (CAGI). The additional experiments indicated that PredPSI-SVR is helpful for prioritizing pathogenic synonymous mutations.

2 | MATERIALS AND METHODS

2.1 | Changes of alternative splicing ($\Delta\Psi$)

The expression level of an alternative exon can be quantified by the fraction of mRNA containing the exon, which is denoted as PSI (Ψ)

$$\Psi = \frac{\text{inclusion reads}}{\text{exclusion reads} + \text{inclusion reads}} * 100\%,$$

where inclusion reads are counts of sequenced fragments aligned to the exon or its junctions with adjacent exons, and exclusion reads are the counts aligned to junctions supporting the exon's exclusion. The inclusion of an exon in the alternative splicing may be affected by genetic variants, especially those occurring around the junction sites. To study the effects of variants on junction sites, the change of Ψ ($\Delta\Psi$) was commonly computed as the differences of Ψ between the wild-type and their variants.

2.2 | Vex-seq data set

All data of variants and their causing $\Delta\Psi$ was downloaded from the CAGI official website (URL: https://genomeinterpretation.org/content/vex-seq). The data set was sequenced by a barcoding approach of variant exon sequencing (Vex-seq), and has been provided by the CAGI 5 organizer to assess methods for the prediction of genomic variants affecting exon splicing. The data set consists of 957 variants distributed on the chromosomes 1 to 8 for model training, namely TR957, and 1,098 variants on the chromosomes 9 to X for the test, namely TS1098. Each variant locates in either a central exon or the flanking intronic region. The CAGI competition is a blind test, where the experimental results of $\Delta\Psi$ in the test set were released after all predictions have been submitted by participants. Therefore, TS1098 is a strictly independent test set for our method.

2.3 | Features

All variants in the Vex-seq data set were annotated by ANNOVAR (Wang, Li, & Hakonarson, 2010) to determine their locations in exons. For each exon, genome sequence was fetched to cover the exon and its flanking regions of 300 nt up- and downstream, from which 42 features were extracted including six splice site motif features, eight splicing regulatory elements, two pre-mRNA secondary structures, Ψ, 17 CADD annotations, $\text{SPIDEX}_{\Delta\Psi}$, and seven features for variants location or codon (Detailed in Table S1). In short, the splice site motif features were calculated by MaxEntScan (Yeo & Burge, 2004), which was applied to scoring 5′ splicing site and 3′splicing site in the WT and mutant (MT) sequences, respectively. These scores were denoted as MES_{5WT}, MES_{3WT}, MES_{5MT}, and MES_{3MT}, respectively. The differences of MES between the MT and the WT sequences were derived from 5′ and 3′ splicing sites and denoted as ΔMES_5 and ΔMES_3.

The splicing regulatory elements used in our models include ESE SR-protein SF2/ASF from ESEfinder (Smith et al., 2006), ESS FAS-hex3 hexamer from FAS-ESS (Wang et al., 2004), and putative ESE and ESS pESE/pESS (Zhang, Kangsamaksin, Chao, Banerjee, & Chasin, 2005). These features were scored using scripts provided by SilVA program (Buske, Manickaraj, Mital, Ray, & Brudno, 2013). As SilVA was designed for only synonymous mutations, we slightly modified

the scripts so that they can be applied to other single-nucleotide variants (SNVs) or indels, in exons or introns.

Pre-mRNA secondary structure features include change of free energy ($\Delta\Delta G$) calculated by UNAfold 3.8 (Markham & Zuker, 2008) and ensemble diversity change (ΔD) calculated by ViennaRNA 2.4.4 (Lorenz et al., 2011).

The Ψ values used in the CAGI challenge data set were provided by the CAGI organizers. To be able to make predictions on datasets without available Ψ values, we calculated Ψ for other exons by using the program MISO 0.5.4 (Katz et al., 2010) over the Human BodyMap 2.0 project (NCBI GSE30611). As the Human BodyMap 2.0 project provides raw RNA-seq data across 16 human tissues, we have separately computed Ψ for each tissue sample to avoid biases to tissues. The alignment of paired-end reads to the reference genome (hg19) was performed by BWA-MEM 0.7.17 (Li & Durbin, 2009). Since the RNA-seq data has a low sequencing depth, MISO cannot obtain Ψ for most exons, and we calculated the average over available tissues for each exon. Finally, we obtained Ψ of 33922 exons, covering about 2.7% of all exons according to the GENCODE GRCh37 gene annotation file (Frankish et al., 2019).

CADD annotation features were extracted from the annotation for variants by CADD (Kircher et al., 2014), which contain DNA conservation scores, histone methylation levels and other sequence features. $\text{SPIDEX}_{\Delta\Psi}$ is a precomputed $\Delta\Psi$ score provided by the SPANR (Xiong et al., 2015). The missing values from SPIDEX were simply filled with zero. We also defined another seven features to describe variants' locations and whether they will introduce stop codons, but they were not chosen by the feature selection.

2.4 | Support vector regression

We trained regression models by the support vector regression implemented in the LIBSVM 3.23 package (Chang & Lin, 2011). LIBSVM is a user-friendly SVM package designed for training SVM model as well as feature scaling, hyperparameter tuning, etc. In this study, ε-SVR in the package and radial basis function (RBF) were selected as the kernel function for the SVM regression. This model has two hyperparameters: the cost parameter C and the gamma (γ) of RBF kernel. To find the best value combination of these two hyperparameters, we adopted a grid search strategy that tests on each combination of $C \in \{2^{-5}, 2^{-3}, \cdots, 2^{11}\}$ and $\gamma \in \{2^{-11}, 2^{-9}, \cdots, 2^{3}\}$.

2.5 | Feature selection

A greedy feature selection algorithm was used as in the previous study (Zhao et al., 2013). In the selection process, we selected the first feature with the highest PCC and used it as the first optimal subset of features. Based on the optimal subset, we scanned all remaining features by adding them individually, and added to the optimal subset with the feature that could mostly improve the accuracy of predicting results. This continued until there was no more feature that could increase performance. During this procedure, we used a 10-fold cross-validation (CV) strategy to evaluate the performance of models, where all variants were randomly separated into 10-folds. Here, variants from the same gene were put into the same fold to avoid sharing gene information between the training and validation sets (Zhao et al., 2018). Every time, nine folds were used for training, and the left fold was used for prediction. This process was repeated for 10 times, and all prediction results were collected to calculate the PCC between predicted $\Delta\Psi$ with experimental values.

2.6 | Synonymous mutation datasets

The change on the alternative splicing was found to be one important factor for pathogenic synonymous mutations (Livingstone et al., 2017). To further test our model and evaluate the relationship between changes on alternative splicing and diseases, we compiled a synonymous mutation data set consisting of both pathogenic and normal mutations. The pathogenic synonymous mutations were downloaded from dbDSM (Wen, Xiao, & Xia, 2016), which is a database for deleterious synonymous mutations collected from public databases and literatures. We first removed duplicate and invalid records, and then converted the chromosome annotation from hg38 to hg19 assembly by using the CrossMap (Zhao et al., 2014). The normal synonymous mutations were obtained from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) with an allele frequency ranging between 0.1 and 0.9. We further removed mutations that are more than 300 bp away from the nearest splice site, leading to 890 pathogenic and 14030 normal synonymous mutations. By applying SPANR and MaxEntScan on these synonymous mutations, there were 133 pathogenic and 3,208 normal mutations having no SPANR score and we removed these mutations from the data set. Finally, 757 pathogenic and 10822 normal mutations remained, namely SynonMut-complete.

Since our method PredPSI-SVR requires an input of Ψ, we mapped the synonymous mutations to exons. After excluding mutations in the exons having no experimental Ψ, we obtained a subset consisting of 87 pathogenic and 826 normal synonymous mutations, namely SynonMut-psi. This data set is 8.7 and 13.1 times smaller than SynonMut-complete in the pathogenic and normal mutations, respectively.

3 | RESULTS AND CONCLUSION

3.1 | Feature analysis

We first computed the PCC between individual features and PSI change ($\Delta\Psi$). As PCC ranges from −1 to 1 with a negative PCC value indicating negative correlation, features were sorted according to the absolute value of PCC values. Table 1 listed nine most important features with an absolute value of PCC greater than 0.1 in the training set TR957. MES scores are a group of most relevant features with five types of MES score features being in the list. ΔMES_5 was the most correlated feature, and ΔMES_3, $\text{MES}_{3\text{MT}}$, $\text{MES}_{5\text{MT}}$, and $\text{MES}_{3\text{WT}}$ ranked the 3th, 4th, 6th, and 8th respectively. MES score was designed to reflect the strength of splice site junction, where a lower

TABLE 1 Top features with the greatest absolute values of Pearson correlation coefficients (PCC) to the ΔΨ computed in the training set TR957

Rank	Feature	PCC (TR957)	PCC (TS1098)
1	**ΔMES_5**	0.402	0.348
2	**$SPIDEX_{\Delta\Psi}$**	0.270	0.241
3	**ΔMES_3**	0.263	0.425
4	**MES_{3MT}**	0.186	0.152
5	dist-Splice	0.176	0.167
6	**MES_{5MT}**	0.167	0.084
7	verPhyloP	−0.128	−0.080
8	**MES_{3WT}**	0.111	−0.019
9	GC	0.108	0.030

Note. Features with absolute values greater than 0.1 are listed. Their PCCs in the test set TS1098 are listed in the last column.
Abbreviations: GC, guanine-cytosine; PCC, Pearson correlation coefficient.

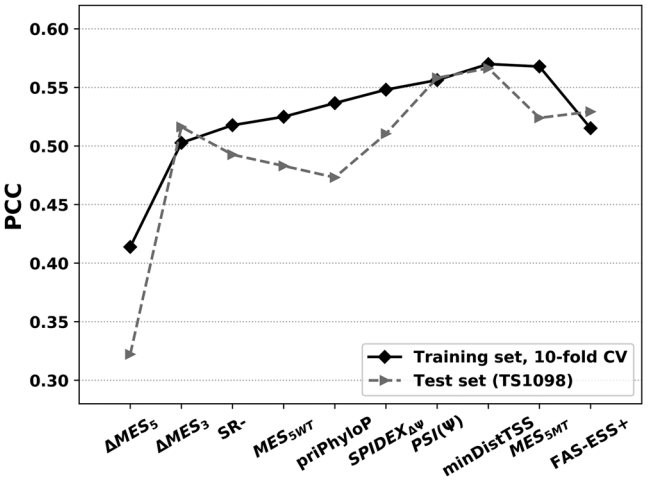

FIGURE 1 The growth of PCC as the number of features increases. The solid line shows the results of 10-fold cross-validation on the training set, and the dashed line for the independent test set. CV, cross-validation; PCC, Pearson correlation coefficient

MES score indicates that the exon is more likely to be interfered by splicing variants (Eng et al., 2004). The second most important feature was $SPIDEX_{\Delta\Psi}$, originally developed for predicting ΔΨ in the previous study (Xiong et al., 2015), which were obtained from the prescored database of the ANNOVAR package (Wang et al., 2010). The fifth important feature "dist-Splice" was the distance of the variant separated from the nearest splice site (5′ or 3′ site). The seventh feature verPhyloP was the phyloP conservation score for vertebrate animals. The last one was GC, which stands for percent GC (guanine-cytosine) in a window size of 75 bp. GC and the verPhyloP were extracted from annotations of CADD. Since pre-mRNA structure was reported to affect splicing (Lin, Taggart, & Fairbrother, 2016), we have evaluated both the free energy change and structural ensemble diversity change measured by both UNAFold 3.8 and ViennaRNA 2.4.4, but they showed only weak correlations with ΔΨ, with the highest PCC of 0.032 by the free energy log change(ΔΔG) computed from the UNAfold 3.8.

3.2 | Model training and feature selection

We have used a greedy feature selection algorithm to select effective features from all 42 features by using 10-fold cross-validation over the training set. As shown in Figure 1, the PCCs by the 10-fold CV gradually increase with the addition of features and reach the highest value of 0.570 by eight features. Further addition of features decreased the performance. In the independent test set, the input of eight features consistently gave the highest PCC, though there is a slight drop in PCCs with three to five features. The most important feature is the ΔMES_5 that individually shows the highest correlation (PCC = 0.402) with the ΔΨ. The other two scores (ΔMES_3, $SPIDEX_{\Delta\Psi}$) gave strong correlation with ΔΨ individually, and the remained five features (ESE feature SR protein loss [SR-], MES_{5WT}, conservation score feature priPhyloP, WT Ψ, and minDistTSS [distance to closest transcript start]) were individually indicated weak correlations with ΔΨ. As shown in Table 2, the combination of these five features and $SPIDEX_{\Delta\Psi}$ can increase the PCC of model predictions from 0.503 to 0.570. In the independent test, the PCC increases from 0.322 by combining two features from MaxEntScan to 0.516, and to 0.566 by (PredPSI-SVR) combining eight features. At the same time, the removal of individual features consistently shows a decrease of PCC, with the largest drop from ΔMES_5, and the smallest from the $SPIDEX_{\Delta\Psi}$. This is probably because the information of $SPIDEX_{\Delta\Psi}$ has been partially covered by other features. Figure 2 shows a comparison between experimental ΔΨ and the predicted ΔΨ by PredPSI-SVR and $SPIDEX_{\Delta\Psi}$. Surprisingly, when we prepared the final server version, we found the removal of priPhyloP and minDistTSS obtained from the CADD leads to slight increase in PCCs of both the 10-fold CV and independent tests compared to the full model with eight features: increase from 0.570 to 0.590 in the 10-fold CV, and from 0.566 to 0.577 in independent test. This indicates the limit of our current greedy

TABLE 2 Performances of models by incremental addition of features, or by removing each feature from the final model tested on the training data set (10-fold cross-validation)

Features added[a]	PCC	Feature excluded[b]	PCC
		Final model	0.570
ΔMES_5	0.414	**-ΔMES_5**	0.444
+ΔMES_3	0.503	**-ΔMES_3**	0.482
+SR-	0.518	**-SR-**	0.555
+MES_{5WT}	0.524	**-MES_{5WT}**	0.542
+priPhyloP	0.537	-priPhyloP	0.545
+$SPIDEX_{\Delta\Psi}$	0.548	**-$SPIDEX_{\Delta\Psi}$**	0.565
+Ψ	0.556	**-Ψ**	0.508
+minDistTSS	0.570	-minDistTSS	0.556

[a]Performance by incremental addition of each feature.
[b]Performance by removing each feature from the final model.

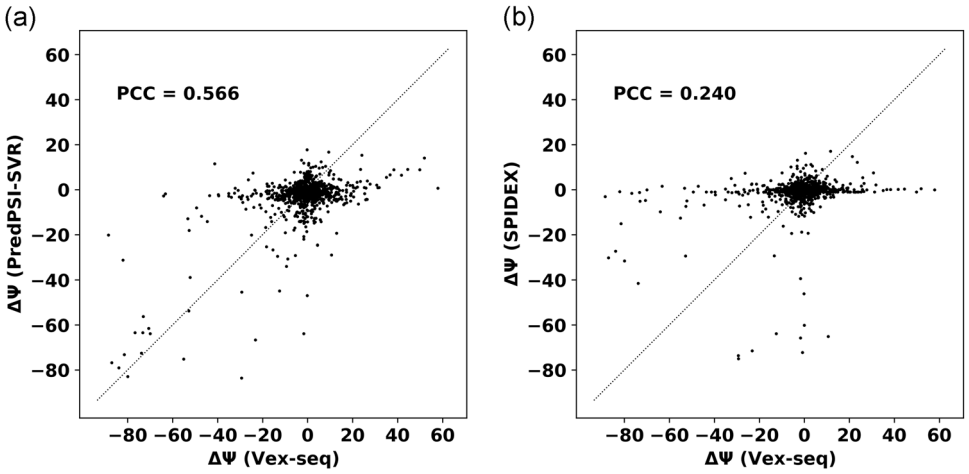

FIGURE 2 Comparison of predicted **ΔΨ** by (a) PredPSI-SVR and (b) **SPIDEX$_{\Delta\Psi}$** (SPANR method) and experimental values on the independent test set TS1098. PCC, Pearson correlation coefficient

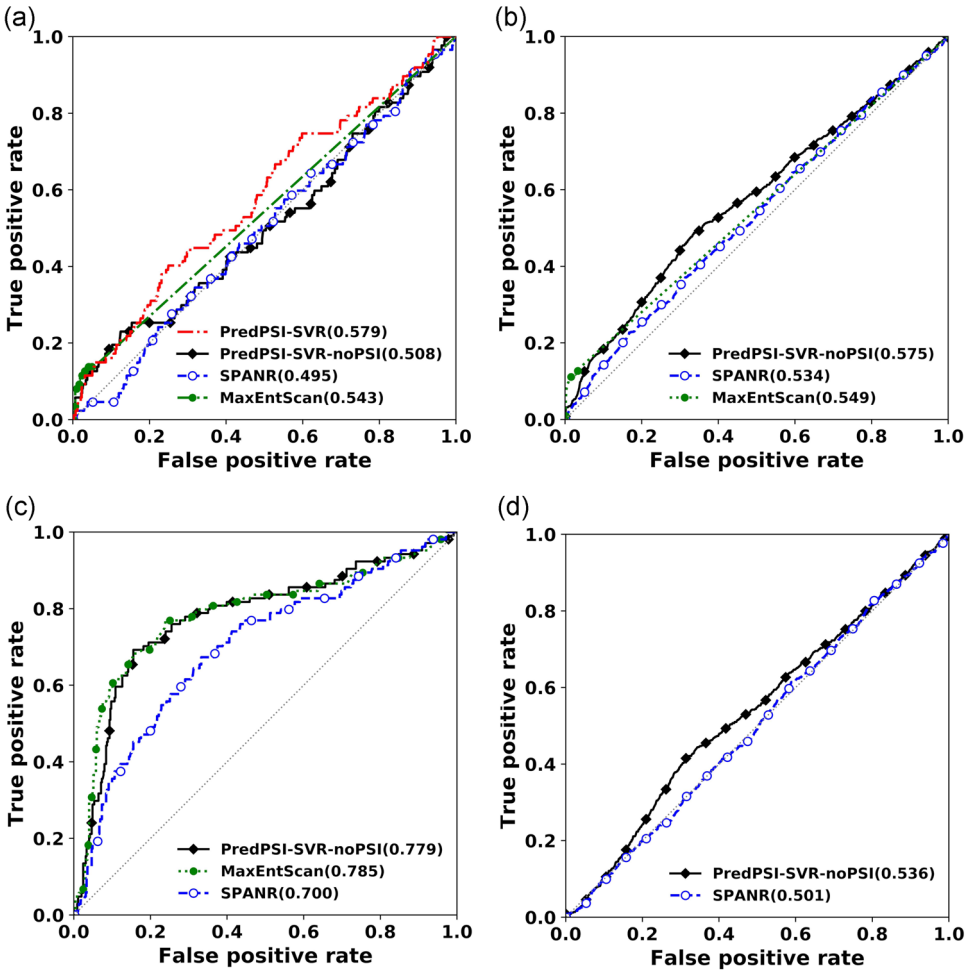

FIGURE 3 ROC curves for PredPSI-SVR, PredPSI-SVR-noPSI, SPANR an MaxEntScan on (a) SynonMut-PSI data set and (b) SynonMut-complete data set. PredPSI-SVR does not appear in the 2nd plot because the data set consists of mutations on exons without experimental PSI values. ROC curves for different methods on mutations (c) with MaxEntScan scores or (d) without MaxEntScan scores were also shown. ROC, receiver operating characteristic curves

feature selection algorithm. Therefore, our final server version (PredPSI-SVR) was trained by using six features over a combination of training and test sets from the CAGI.

The CAGI provides experimental Ψ for a small portion of exons, and MISO cannot compute Ψ for all exons. For general use where exons do not have Ψ values, we have built another model, PredPSI-SVR-noPSI without using the Ψ. The model achieved lower performance with PCCs of 0.525 and 0.479 on the 10-fold CV of the training set and the independent test set, respectively.

3.3 | The prioritization of pathogenic synonymous mutations

The PredPSI-SVR model was further utilized to prioritize pathogenic synonymous mutations, and compared with SPANR and MaxEntScan. For PredPSI-SVR and SPANR, we directly used the absolute values of predicted ΔΨ. For MaxEntScan, we took the sum of the absolute values of ΔMES_5 and ΔMES_3 to obtain information for both 5′ and 3′ splicing sites. These scores were used to distinguish pathogenic mutations from normal ones. Mutations with a score above a threshold will be classified as pathogenic. As shown in Figure 3, we plotted the receiver operating characteristic curves (ROC) by the PredPSI-SVR, SPANR, and MaxEntScan methods on the SynonMut-PSI data set. PredPSI-SVR performs the best, whereas SPANR performs the worst that is close to random on the data set. As shown in Table 3, the area under ROC (AUC) indicates that PredPSI-SVR is significantly better than SPANR (p value = .036), and 6.6% higher than MaxEntScan. The PredPSI-SVR-noPSI without an input of experimental Ψ has a big drop in the AUC (from 0.579 to 0.508) likely due to the small data set. On the larger SynonMut-complete dataset, PredPSI-SVR-noPSI achieves an AUC of 0.575, which is significantly better than the SPANR and MaxEntScan with p values of .004 and .049, respectively according to the statistical test (Hanley & McNeil, 1982). The Hanley & McNeil test is a statistical method for testing whether there is a significant difference between two AUC values. These results also indicate that our predictions on changes of alternative splicing can help in prioritizing pathogenic synonymous mutations. At the region of low FPR (FPR < 0.1), the curve of MaxEntScan is slightly above the one for PredPSI-SVR though MaxEntScan is an input feature for the PredPSI-SVR model. This is likely because our method was optimized for the overall performance that has brought down the results in this region. The problem may be overcome by using other machine learning algorithms like XGBoost (Chen & Guestrin, 2016) or a bigger training data set. In addition, we divided the SynonMut-complete data set into two portions: 555 mutations within the scanning scope of MaxEntScan and the remaining 11,024 mutations. For the first portion, our model has essentially the same performance as MaxEntScan (Figure 3c), while for the remaining mutations without MaxEntScan scores, our model achieves an AUC of 0.536 that is significantly better than the AUC (=0.501) by SPANR with a p value of .018 (Figure 3d). These suggest that our model can utilize additional features in addition to the MaxEntScan scores.

TABLE 3 The performance of methods to discriminate pathogenic from normal synonymous mutations

Data set	Methods	AUC	p value[a]
SynonMut-PSI	PredPSI-SVR	0.579	–
	PredPSI-SVR-noPSI	0.508	– (.064)[b]
	SPANR	0.495	0.389 (.036)
	MaxEntScan	0.543	0.150 (.220)
SynonMut-complete	PredPSI-SVR-noPSI	0.575	–
	SPANR	0.534	.004
	MaxEntScan	0.549	.049

Abbreviations: AUC, area under ROC; ROC, receiver operating characteristic curves.

[a]The significance of difference between methods compared to PredPSI-SVR-noPSI.

[b]PredPSI-SVR (values in the parenthesis) according to the statistical test (Hanley & McNeil, 1982).

4 | DISCUSSION

In this study, we present a new method, namely PredPSI-SVR to predict the change of exon splicing caused by genetic variants. PredPSI-SVR is a support vector regression model that integrates features of splice sites, splicing regulatory elements, DNA conservation score, $SPIDEX_{\Delta\Psi}$ provided by SPANR, and Ψ of WT exons to predict ΔΨ. The method achieved PCCs of 0.570 and 0.566 for the 10-fold CV on the training data set and strictly independent test set, respectively. This performance is significantly better than the performance (PCC = 0.24) by SPANR's $SPIDEX_{\Delta\Psi}$.

To build such a model, we extracted 42 features at first and analyzed their correlations with ΔΨ. We found that features on splicing sites computed by MaxEntScan have the highest correlations. The model trained by the ΔMES_5 and ΔMES_3 can achieve a PCC of about 0.51 on the test set, indicating the importance of variants around splice sites to affect the alternative splicing. By using greedy feature selection, the model built from eight selected features increased the PCCs from 0.503 to 0.570 on training set and from 0.516 to 0.566 on test set. Five among eight selected features individually shows weak correlations with ΔΨ (|PCC| < 0.1), indicating importance to extract comprehensive features.

Our method ranked the 2nd in the CAGI challenge, and it is of interest to compare with other methods. According to the descriptions of the prediction methods, available at https://genomeinterpretation.org/content/vex-seq, two groups (groups 1 and 2, which were ranked 3th and 4th, respectively) used similar features to our method (group 4). In contrast to our approach, the group 1 did not fit their model directly toward the experimental ΔΨ values. They trained a classification model to predict the sign of ΔΨ and then used the predicted scores to fit into the ΔΨ. The group 2 did not employ a cross-validation to optimize the hyperparameters for their random forest model, which might cause a lack

of generalization to the test set. The group 3 (ranked 5th) did not provide implementation details. On the other side, the group 5 made the best predictions by using their developed MMSplice method (Cheng et al., 2019). In their method, six deep neural networks have been trained to extract features of splice donor, splice acceptor, 5′ exon, 3′ exon, 5′intron, and 3′ intron, which were later combined by a simple linear regression to predict ΔΨ. With the benefit of utilizing deep-learning techniques, the method achieved a PCC of 0.675 for the test set. Therefore, the predictions might be further improved by coupling merits of different methods, for example, using features mined from deep learning combined with our used knowledge-based information like conservation scores, and training a nonlinear model with machine learning methods like SVM, as used in our method.

We also noticed that the removal of mutations with small Ψ changes lead to a better correlation between the predicted and experimental values. By removing ΔΨ with absolute values less than two times of the standard deviation, increasing correlations were observed for all methods on the remained 53 mutations. For example, our PredPSI-SVR achieved an increase in PCC from 0.566 to 0.665 and the top method MMSplice increased from 0.675 to 0.782. This is likely because the mutations with a small change of Ψ might be affected by many other factors with relatively weak impact, while current methods can only capture the dominant factors due to the limited data.

At present, PredPSI-SVR does not include features of branch point sequences and pre-mRNA secondary structure effectively due to the limit by relatively small numbers of samples in the data set. Moreover, the small number of samples prevented us from using more powerful classification algorithms like deep learning. Another limitation of our method is its need for experimental Ψ of exons. Without experimental Ψ, PCCs on the training set and test set dropped by about 0.1. Currently only RNA-seq data from the Human BodyMap Project 2.0 project was used, and many exons cannot be found from the MISO analysis due to sequencing depth. With advances in sequencing technology, an increasing number of public databases are becoming available. This enables us to capture more accurate Ψ for exons, and thus to improve the performance. Moreover, the tissue dependence of Ψ reminds us to use tissue-specific Ψ in PredPSI-SVR to better discover pathogenic variants in specific diseases.

The PredPSI-SVR method is available with a standalone version on https://github.com/chenkenbio/PredPSI-SVR. The program runs on Linux/Unix system with input of variants in the VCF format.

ACKNOWLEDGMENTS

This project was supported in part by the National Natural Science Foundation of China (61772566, U1611261, and 81801132), the program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211) and Guangdong Province Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation (2017B030314026). We would like to thank the Critical Assessment of Genome Interpretation (CAGI) group and data providers. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

ORCID

Ken Chen http://orcid.org/0000-0001-5701-1438

REFERENCES

Adamson, S. I., Zhan, L., & Graveley, B. R. (2018). Vex-seq: High-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biology*, *19*(1), 71. https://doi.org/10.1186/s13059-018-1437-x

Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: Examples, detection, and confirmation. *Journal of Applied Genetics*, *59*(3), 253–268. https://doi.org/10.1007/s13353-018-0444-7

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, *18*, 437–451.

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., & Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics*, *29*(15), 1843–1850. https://doi.org/10.1093/bioinformatics/btt308

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27. 1–27:27

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/.org/10.1145/2939672.2939785

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, ž., & Gagneur, J. (2019). MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, *20*(1), 48. https://doi.org/10.1186/s13059-019-1653-z

Climente-González, H., Porta-Pardo, E., Godzik, A., & Eyras, E. (2017). The functional impact of alternative splicing in cancer. *Cell Reports*, *20*(9), 2215–2226. https://doi.org/10.1016/j.celrep.2017.08.012

Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., & Gatti, R. A. (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths. *Human Mutation*, *23*(1), 67–76. https://doi.org/10.1002/humu.10295

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., & Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Furuya, M., Kobayashi, H., Baba, M., Ito, T., Tanaka, R., & Nakatani, Y. (2018). Splice-site mutation causing partial retention of intron in the FLCN gene in Birt-Hogg-Dubé syndrome: A case report. *BMC Medical Genomics*, *11*(1), 42. https://doi.org/10.1186/s12920-018-0359-5

Goodman, D. B., Church, G. M., & Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science*, *342*(6157), 475–479. https://doi.org/10.1126/science.1241934

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–1015. https://doi.org/10.1038/nmeth.1528

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lin, C.-L., Taggart, A. J., & Fairbrother, W. G. (2016). RNA structure in splicing: An evolutionary perspective. *RNA Biology*, *13*(9), 766–771. https://doi.org/10.1080/15476286.2016.1208893

Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D. N., & Zhou, Y. (2017). Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Human Mutation*, *38*(10), 1336–1347. https://doi.org/10.1002/humu.23283

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*, 26. https://doi.org/10.1186/1748-7188-6-26

Markham, N. R., & Zuker, M. (2008). UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular Biology (Clifton, N.J.)*, *453*, 3–31. https://doi.org/10.1007/978-1-60327-429-6_1

McManus, C. J., & Graveley, B. R. (2011). RNA structure and the mechanisms of alternative splicing. *Current Opinion in Genetics & Development*, *21*(4), 373–379. https://doi.org/10.1016/j.gde.2011.04.001

Parmley, J. L., Chamary, J. V., & Hurst, L. D. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular Biology and Evolution*, *23*(2), 301–309. https://doi.org/10.1093/molbev/msj035

Rosenberg, A. B., Patwardhan, R. P., Shendure, J., & Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, *163*(3), 698–711. https://doi.org/10.1016/j.cell.2015.09.054

Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q., & Krainer, A. R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Human Molecular Genetics*, *15*(16), 2490–2508. https://doi.org/10.1093/hmg/ddl171

Supek, Fran, Miñana, Belén, Valcárcel, Juan, Gabaldón, Toni, & Lehner, Ben (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, *156*(6), 1324–1335. https://doi.org/10.1016/j.cell.2014.01.051

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Wang, G.-S., & Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, *8*(10), 749–761. https://doi.org/10.1038/nrg2164

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, *14*(5), 802–813. https://doi.org/10.1261/rna.876308

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic Identification and Analysis of Exonic Splicing Silencers. *Cell*, *119*(6), 831–845. https://doi.org/10.1016/j.cell.2004.11.010

Wen, P., Xiao, P., & Xia, J. (2016). dbDSM: A manually curated database for deleterious synonymous mutations. *Bioinformatics*, *32*(12), 1914–1916. https://doi.org/10.1093/bioinformatics/btw086

Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, *3*(7), a003707. https://doi.org/10.1101/cshperspect.a003707

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., & Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science*, *347*(6218), 1254806–1254806. https://doi.org/10.1126/science.1254806

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *11*(2–3), 377–394. https://doi.org/10.1089/1066527041410418

Zhang, X. H.-F., Kangsamaksin, T., Chao, M. S. P., Banerjee, J. K., & Chasin, L. A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. *Molecular and Cellular Biology*, *25*(16), 7323–7332. https://doi.org/10.1128/MCB.25.16.7323-7332.2005

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. -P., & Wang, L. (2014). CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

Zhao, H., Yang, Y., Lin, H., Zhang, X., Mort, M., Cooper, D. N., & Zhou, Y. (2013). DDIG-in: Discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biology*, *14*(3), R23. https://doi.org/10.1186/gb-2013-14-3-r23

Zhao, H., Yang, Y., Lu, Y., Mort, M., Cooper, D. N., Zuo, Z., & Zhou, Y. (2018). Quantitative mapping of genetic similarity in human heritable diseases by shared mutations. *Human Mutation*, *39*(2), 292–301. https://doi.org/10.1002/humu.23358

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chen K, Lu Y, Zhao H, Yang Y. Predicting the change of exon splicing caused by genetic variant using support vector regression. *Human Mutation*. 2019;40:1235–1242. https://doi.org/10.1002/humu.23785