IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2017, accepted November 7, 2017, date of publication November 20, 2017,
date of current version December 22, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2775703

Identify Key Sequence Features to Improve CRISPR sgRNA Efficacy

LEI CHEN[1], SHAOPENG WANG[2], YU-HANG ZHANG[3], JIARUI LI[2], ZHI-HAO XING[3], JIALIANG YANG[4], TAO HUANG[3], AND YU-DONG CAI[2]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[2]School of Life Sciences, Shanghai University, Shanghai 200444, China
[3]Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[4]Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA

Corresponding authors: Lei Chen (chen_lei1@163.com); Tao Huang (tohuangtao@126.com); Yu-Dong Cai (cai_yud@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 31371335 and Grant 31701151, in part by Natural Science Foundation of Shanghai under Grant 17ZR1412500, and in part by the Shanghai Sailing Program and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2016245.

ABSTRACT The CRISPR/Cas9 system is a creative and innovative gene editing biotechnology tool in genetic engineering. Although several achievements have been attained using the CRISPR/Cas9 system, it is still a challenge to avoid off-target effects and improve the editing efficacy. Previous efforts on evaluating the efficacy and designing the guide RNA mainly focused on DNA properties. However, some DNA features have not been characterized but can be reflected by protein properties, such as the disorder features and the sequence conservation. In this paper, we provided a computational framework to identify important features related to the efficacy of CRISPR/Cas9 focusing on the properties of the proteins encoded by the target DNA fragments. The feature selection method, maximal-relevance-minimal-redundancy, was adopted to analyze these features. And incremental feature selection together with support vector machine, were employed to extract optimal features, on which an optimal classifier can be constructed. As a result, 152 important features were extracted, with which an optimal classifier based on support vector machine was built. This classifier obtained the highest MCC value of 0.355. Finally, a series of detailed biological analyses were performed on the optimal features. From the results, we found that some key factors may differentially affect the binding activity of sgRNAs to their targets. Among them, the disorder status of the target protein sequences was found to be a major factor that is related to the efficacy of sgRNAs, suggesting the DNA features associated with the protein disorder status could also affect the CRISPR/Cas9 efficacy.

INDEX TERMS CRISPR/Cas9 system, sgRNAs, maximal-relevance-minimal-redundancy, incremental feature selection, protein disorder.

I. INTRODUCTION

A protein-coding gene has been widely regarded as a locus of DNA that can be transcribed into messenger RNA and translated into a polypeptide chain to exert specific biological functions [1], [2]. Generally, different genes have different functions and may play different roles in various biological processes. As our understanding of gene functions are improved, research has aimed at developing functional tools to explore and even alter the specific function of a known gene [3], [4]. However, without functional and applicable molecular biology tools, it is quite difficult to accurately change and regulate the function of a specific gene. Therefore, to precisely control the functions of genes, various biotechnologies have been developed and applied.

Development of gene editing biotechnologies has led to three main subtypes of gene editing technologies: ZFN (zinc-finger nucleases) TALEN, (transcription activator-like effector nuclease) and CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats). The first strategy for genome customization was ZFN [5]. The zinc-finger nuclease strategies contribute to genome editing via various zinc finger modules, which recognize 3-4 base pairs [5], [6]. Another technology, TALEN, has also been confirmed to be efficient in gene editing [7]. Although TALEN can complete gene editing more quickly and is cheaper and better than ZFN, it is still quite difficult and time-consuming for most laboratories in the world [8]. Therefore, to improve the efficiency and reduce costs, a new technology, CRISPR/Cas9 has been developed and presented [9], [10]. Different from TALEN and ZFN, CRISPR/Cas9 technology is a functional RNA-guided Cas9 nuclease-dependent gene editing technology [11]. The CRISPR/Cas9 regulatory mechanism under nature conditions is an acquired immunity mechanism to fight against foreign plasmids and viruses in bacteria and archaebacteria [12], [13]. With a length of more than twenty nucleotides, the CRISPR/Cas9 system greatly improves the recognition and editing accuracy [11]. In addition, CRISPR/Cas9 technology has much shorter test and experimental periods compared to TALEN and ZFN, and it can edit multiple genome sites simultaneously in the same research system [11]. Therefore, CRISPR/Cas9 is a significant technological improvement that has great advantages compared to TALEN and ZFN. Thus, it has attracted wide attention around the world.

The design of the CRISPR/Cas9 system can be separated into three steps: the selection of target genes and sequences, the design of the sgRNA and the transfection of the integrated plasmids [14], [15]. The specificity of the guide RNAs, and the efficacy of the plasmid transfection and expression are two key points in the protocol of this newly developed technology. Based on these two key points, various publications have contributed to the modification of each step of the CRISPR/Cas9 system to improve the design principles [16]–[18]. It has been recently reported that modification of the Cas9 nucleases using structural information may alter the recognition ability of alternative PAM (protospacer adjacent motif) sequences and further improve the efficacy [19]. With the extensive use and modification of this new technology, various off-target effects have been reported and solved, while differences in the recognition and editing efficiency have been partially revealed and identified. Recently, with the assistance of computational technologies, various studies have successfully simulated the CRISPR/Cas9 system and modified the technical processes. Listgarten from the eScience Research Group at Microsoft Research in Los Angeles has presented an *in silico* predictive modeling approach to predict the guide efficiency of the CRISPR/Cas9 system. To build up the optimal predictive modeling approach, they summarized all the detailed information reported on the CRISPR/Cas9 system, which our study mainly relied on. Later, Listgarten and Root, from the Broad Institute of MIT and Harvard, further reported a new computational method to optimize sgRNA design and improve the efficacy of CRISPR/Cas9 system [20]. Although such studies have drastically modified the technical processes, the detailed mechanisms that affect the recognition or the editing efficiency have not been fully revealed [14], [15]. Efforts have been made to identify the efficacy related DNA features such as the GC content, binding specificity, alignment identity, amino acid cut position, and amino acid composition of the peptides encoded [20], [21]. However, some DNA features were not investigated by these studies. These DNA features remained uncharacterized but can reflected by the properties of encoded proteins. For example, the protein disorder status were suggested to be associated with the codon usage [22], indicating the specific connection between the DNA features with certain protein properties.

In this study, we identified CRISPR sgRNA efficacy by using protein sequence features translated from the target DNA for the first time. Based on the data from http://research.microsoft.com/en-us/projets/azimuth, we obtained detailed information reported on the CRISPR/Cas9 system [20]. Doench *et al.* provided a resource for the design of improved sgRNA reagents for large-scale screens and gene editing experiments, and they developed metrics to predict off-target sites and effects [20]. However, the important factors and underlying mechanisms that contribute to the observed variable gene editing efficacy and accuracy have not been fully elucidated. To identify the core factors that affect the efficiency and accuracy of the CRISPR/Cas9 system, we classified candidate associated factors into three main groups incorporating two protein properties: (1) the stability and variability of the target proteins (which is described as the sequence disorder status); (2) the evolutionary conservation (which is described by the position-specific scoring matrix, PSSM); and (3) the nucleotide composition of the sgRNA. Based on several reliable computation methods, including maximal-relevance-minimal-redundancy (mRMR) [23], incremental feature selection (IFS), support vector machine (SVM) [24], [25], we presented a new computational framework to screen a group of core factors that may affect the efficiency of the CRISPR/Cas9 system. Results yielded by this framework may help us deepen our understanding of this new and important biotechnology.

II. MATERIAL AND METHODS

A. DATASET

To predict the activity of sgRNAs on target proteins, a reliable and qualified dataset needed to be constructed. We downloaded the sgRNA efficacy data, which was reported in Doench *et al.*'s study [20], from http://research.microsoft.com/en-us/projets/azimuth. The associated description of the 17 target genes and the corresponding lengths of the 17 protein sequences are listed in **Table 1**. In the original dataset containing 5,310 sgRNA-target pairs, four items were selected to represent each pair: (I) the target gene; (II) the cut position in the protein translated from the DNA cut position; (III) the 30mer sgRNA in the target site; and (IV) the efficacy score. The efficacy scores of the specific sgRNA-target pairs represent the activity or binding affinity of the sgRNAs.

To construct a well-defined dataset, a data cleaning procedure was executed on the aforementioned dataset. First, if more than one sgRNA-target pairs with slightly different efficacy scores cut in same position, then one sgRNA-target pair was randomly selected and was used to represent that specific cut position, resulting in 3,345 sgRNA-target pairs for further utilization. Next, a 21-amino acid

TABLE 1. The detailed description of the 17 target genes and proteins.

Gene name	Length of protein sequence (aa)	CCDS number	Number of positive samples	Number of negative samples
CCDC101	293	10635.1	48	47
CD13	967	10356.1	156	142
CD15	530	8301.1	97	92
CD28	220	2361.1	17	13
CD33	364	33084.1	56	48
CD43	400	10650.1	9	6
CD45	1306	1397.2	4	4
CD5	495	8000.1	35	44
CUL3	768	2462.1	60	65
H2-K1	369	50069.1	65	63
HPRT1	218	14641.1	22	24
MED12	2177	43970.1	332	323
NF1	2839	42292.1	299	301
NF2	595	13861.1	86	84
TADA1	335	1255.1	48	37
TADA2B	420	47007.1	74	68
THY1	161	8424.1	14	16

peptide segment representing each cut position by combining ten residues in its upstream and downstream, respectively. Then, the 30mer sgRNA was aligned with the corresponding 21-amino acid peptide at the same cut position using BLASTX [26] and the target protein was obtained if a successful alignment occurred. The resulting dataset consisted of 2,799 sgRNA-target pairs that successfully aligned with the target protein. At the same time, the identity of the target protein was appended onto the information for each sgRNA-target pair to represent the activity of sgRNA for each cut position.

In this study, we tried to use some advanced machine learning algorithms to extract important factors that can influence the efficacy of sgRNAs. Thus, all 2,799 sgRNA-target pairs were divided into two partitions according to their efficacy scores. sgRNA-target pairs with efficacy scores greater than 0.5 constituted one partition and were regarded as "positive samples", while the rest pairs, i.e., those with less than or equal to 0.5, comprised the other partition and were called "negative samples". Accordingly, we constructed a dataset composing of 1,377 negative samples and 1,422 positive samples. In this way, the problem of predicting the sgRNA activities were transformed into a binary classification problem. The main purpose of this study was to extract important factors that give important contribution for this classification problem. The description of the 1,377 negative and 1,422 positive samples is listed in Table S1.

B. FEATURE CONSTRUCTION

As described in Section II.A, the activity of sgRNA at each cut position was represented by its related 30mer sgRNA and its target protein. Given this information, one type of features was derived from the 30mer sgRNA, and two types of features were derived from the target protein; these three types of features were used to represent each sgRNA-target pair. These features included: (1) single and pair-wise nucleotides (SNTs and PNTs); (2) position specific scoring matrix (PSSM); and (3) disorder status, all of which have been widely used in several studies [27]–[32].

1) SNT AND PNT FEATURES

Following the one-letter code of four types of nucleotides (A, C, G and T), each 30mer sgRNA was encoded as a four-dimensional vector. In these vectors, each position was set to one or zero depending on the identity of the base. For example, a "G" in sgRNA was encoded as a vector of "0010". Likewise, the pair-wise nucleotides in the position of $[i, i+1]$ ($i = 1, 2,.., 29$) were encoded as a sixteen-dimensional vector according to the possible pairwise combinations of all single nucleotides (AA, AC, AG,….., TG, TT). For example, a "CG" in sgRNA was encoded as a vector of "0000001000000000", in which the 7th position was set to one. Accordingly, each 30mer sgRNA was represented by 584 (30×4+29×16) SNT and PNT features.

2) PSSM FEATURES

The sequence conservation could be another factor affecting the efficiency of CRISPR/Cas9. Here we investigated the conservation in the translated protein level as all sgRNA in the dataset that were designed to target the CDS of protein-coding genes [20], which in some extent reflect the DNA conservation. The technique of position specific iteration BLAST (PSI-BLAST) was used to search for remote homologous proteins against the query protein. For a query protein sequence, each amino acid residue is represented by twenty values, which indicate the mutation frequency of the twenty native amino acids against the given residue. For each target protein, the PSI-BLAST [33] was utilized to retrieve the UniRef100 (Release: 15.10 03-Nov-2009) database with three iterations and cutoff E-value of 0.0001. If the length of a given target protein was shorter than ten, one or more "X's" denoting the empty amino acid residues were appended to the C-terminus of the peptide segment. The empty residue was represented by an empty vector with twenty zero. Accordingly, a 10-amino acid target protein was encoded as 200 (10×20) PSSM features.

3) DISORDER STATUS FEATURES

As we mentioned above, protein disorder is associated with codon usage [22], which provides us another angle to address the CRISPR/Cas9 efficacy evaluation. The VLS2 program [34], using the protein sequence as the input, was utilized to calculate the probability of the given residues existing in disordered regions. The output score for each amino acid residue, ranging from zero to one, was denoted as its disorder status, and a higher score indicated a greater possibility of the given residue existing in a disordered region. Accordingly, 10 disorder status features were used to represent the disorder property of the residues in the target proteins.

TABLE 2. Three types of features used to encode the sgRNAs-target pairs.

Feature type	Description	Number of features
SNT and PNT	Single and pair-wise order of nucleotides	584
PSSM	Evolutionary conservation of residues	200
Disorder status	Disorder of residues	10
Total	---	794

In total, 794 (584+200+10) features can be derived from the nucleotide and peptide segments, and these features were utilized to encode each sgRNA-target pair in the dataset. The distribution of these 794 features is listed in **Table 2**. Additionally, the detailed description of 794 features is provided in Table S2. These features would be analyzed by some feature selection methods and key features that may influence the efficiency of CRISPR/Cas9 would be extracted. Compared to other sequence analysis studies from the point view of biological mechanism, features mentioned above were widely applied for the purpose of designing computational methods. Analysis on these features may provide new insights for better comprehension of CRISPR/Cas9.

C. FEATURE SELECTION

As described in Section II.B, a total of 794 features were used to represent the sequence features that may influence the activity of the sgRNAs. It is clear that not all of the 794 features contribute equally in this regard. A feature selection procedure was necessary to extract key features among them [28], [30]–[32], [35]–[45]. Thus, the mRMR method [23] was utilized to rank the 794 features, which would be further used for selection of key features.

During the classification, a target class was assigned to each sample. According to the relationships between the features and their target classes, each feature was ranked in descending order; the resulting list was called the MaxRel feature list. However, it has been found that the usage of some top individual features in the MaxRel feature list does not always lead to good predictions due to the redundancy among these features. Thus, the mRMR method also ranked features according to not only their relevance to the target class but also the redundancy of features in another feature list, called the mRMR feature list. The brief description of yielding this feature list is introduced below. Initially, the mutual information (MI) is calculated to measure the relevance between two variables x and y, which was defined in **Eq. 1**:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy, \tag{1}$$

where $p(x)$ and $p(y)$ represent the marginal probabilistic density of variables x and y and $p(x, y)$ denotes their joint probabilistic density.

In the measurement of maximal relevance, the selected feature f is required to have the largest value of MI to the target class c. The MI value, denoted as D, between feature f and target class c is shown in **Eq. 2**:

$$D = I(f, c) \tag{2}$$

Let Ω be the feature set with N features, Ω_s be a feature set containing the selected features and Ω_t includes the rest features, i.e., $\Omega_t = \Omega$ - Ω_s. The relevance R between a given feature f in Ω_t and all features in Ω_s is defined in **Eq. 3**:

$$R = \frac{1}{|\Omega_S|} \sum_{f_i \in \Omega_S} I(f, f_i) \quad (R = 0 \text{ if } \Omega_s \text{ is empty}) \tag{3}$$

For each feature f in Ω_t, the value D-R is calculated, and the feature with the maximal D-R value is removed from Ω_t and put into Ω_s. When all features are in Ω_s, the whole procedures stop.

According to the selection order of each feature, the mRMR feature list can be constructed in a way that the first selected feature receives the top place in the list, the second selected feature gets the second place, and so forth. The mRMR feature list is illustrated in **Eq. 4**:

$$S = [f_1, f_2, \ldots, f_N] \tag{4}$$

To select important features from the mRMR feature list, the IFS method constructs a series of feature sets, say $S_1, S_2, \ldots, S_N$, such that $S_i = \{f_1, f_2, \ldots, f_i\}$. For each feature set, a classification algorithm was executed on a dataset, in which each sample is represented by features in the set. The feature set yielding the best performance can be extracted. Features in this set were deemed as optimal features and the classifier based on these features is called the optimal classifier.

D. CLASSIFICATION ALGORITHM

As described in Section II.C, the IFS method creates a series of feature sets. To test the discriminating power of each feature set and select the best one, a proper classification algorithm should be adopted. In this study, the classic machine learning algorithm, SVM [24], [25], was employed. Its brief introduction is as below.

SVM [24], [25] is applied to investigate the problems of pattern recognition, regression and classification based on statistical learning theory, which is especially applicable for the modeling of small datasets [46]–[48]. The basic principle of SVM is to map linear non-separable data in low-dimensional space to high-dimensional space so that the mapped data in high dimension can be optimized and separated by a hyper-plane. A query sample is then mapped into the same high-dimensional space, and its predicted class is determined according to which side of the hyper-plane the sample belongs to. In this study, the algorithm of sequential minimal optimization [49] was utilized to train the SVM classifier. In the training process, the modeling problem was split into a series of the smallest possible linearly related sub-problems. Then, the sub-solutions of the sub-problems were combined as the final solution of the original modeling problem. In Weka [50], the classifier, SMO, implements this type of SVM and was directly used in this study. The kernel function was the polynomial.

E. PREDICTION PERFORMANCE MEASUREMENT

To validate the effectiveness of constructed classifiers, some widely accepted and reliable measurements were necessary. Four measurements called sensitivity (*SN*), specificity (*SP*), accuracy (*ACC*) and Matthew's correlation coefficient (*MCC*) [51] were calculated to evaluate the predicted ability of all classifiers. Their definitions are listed in **Eq. 5** to **Eq. 8**.

$$SN = \frac{TP}{TP + FN}, \tag{5}$$

$$SP = \frac{TN}{TN + FP}, \tag{6}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{8}$$

where *TP* (true positive), *TN* (true negative), *FP* (false positive) and *FN* (false negative) can be directly obtained by comparing the target classes and predicted classes of samples in the dataset. The descriptions of the four values are shown as below:

TP: the number of positive samples predicted correctly;

TN: the number of negative samples predicted correctly;

FP: the number of negative samples predicted incorrectly;

FN: the number of positive samples predicted incorrectly.

Among four measurements listed in **Eqs. 5-8**, the *MCC* is a balanced measurement even if the dataset is great unbalanced. It has been used as a major measurement in several studies [52]–[56]. Accordingly, the *MCC* was also selected as the major measurement to validate the classifiers in this study.

III. RESULTS

As introduced in Section II.C, the mRMR method was utilized to rank all of the 794 features. Features were first ranked by their maximal relevance to the target class, resulting in the MaxRel feature list. Then, they were sorted by their maximal relevance to the target class and minimal redundancy to the already selected features, which produced the mRMR feature list. These two lists are provided in Tables S3 and S4.

From the mRMR feature list, a series of feature sets were constructed using the IFS method. For each feature set, The SVM was executed on the dataset, in which each sample was represented by features in the set, with its performance evaluated by 10-fold cross-validation [57]–[59]. The predicted results were counted as *SN*, *SP*, *ACC* and *MCC*. For easy observation, an IFS-curve was plotted using *MCC* as its Y-axis and the number of used features as its X-axis, as shown in **Figure 1**. The best *MCC* 0.355 was obtained when the feature number was pointed to 152. Thus, the optimal classifier based on SVM used the top 152 features in the mRMR feature list to represent sgRNA-target pairs. The performance of the optimal classifier evaluated by the four measurements is shown in **Table 3**. Therefore, the top 152 features in the mRMR feature list are deemed to be the optimal features for identification of highly active sgRNAs. The following section would give detailed analyses on them.

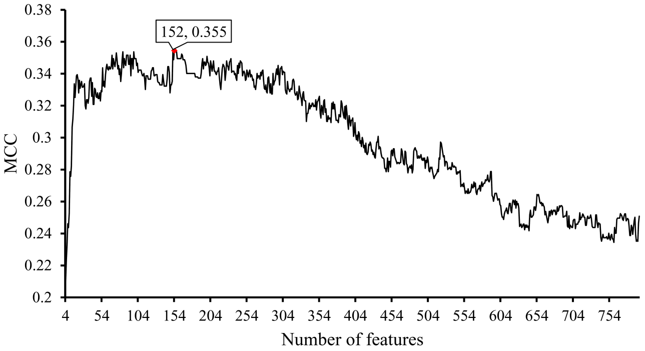


FIGURE 1. The IFS-curve. The X-axis denotes the number of features participating in the construction of classifiers and the Y-axis denotes the corresponding MCC values yielded by SVM.

TABLE 3. The performances of the optimal classifier derived from SVM.

Number of optimal features	SN	SP	ACC	MCC
152	0.698	0.657	0.678	0.355

IV. DISCUSSION

The mRMR method produced the MaxRel feature list, in which features were ranked according to their relevance to target variable. Here, top 10% (80) features in this list were extracted for investigating which feature types are more important. Among the 80 features, 16 of these features belonged to the SNT and PNT, 54 features were derived from the PSSM, and 10 features were features of disorder status. Because the number of features in each type was

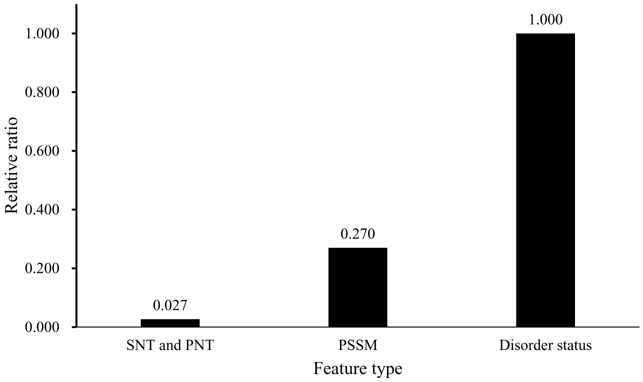


FIGURE 2. The relative ratio for each type of features derived from the top 80 features in MaxRel feature list.

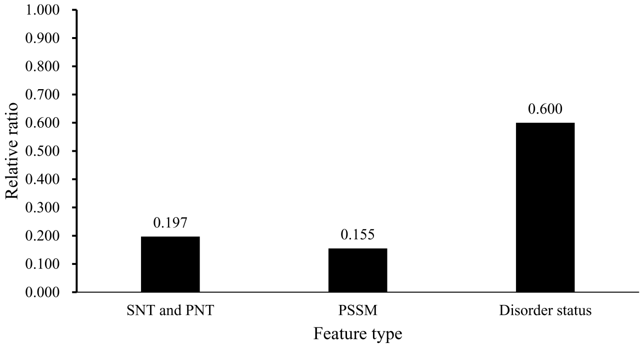


FIGURE 3. The relative ratio for each type of features derived from the top 152 features in mRMR feature list.

not same, only considering the absolute value of number of selected features cannot correctly measure the contribution of each feature type. For example, there were 584 features for SNT and PNT, 16 were included in the top 10% of the MaxRel feature list. However, it may not be more important than disorder status features because all ten disorder status features were included in the top 10% of the MaxRel feature list and top nine features are all disorder status. In view of this, the ratio of each feature type appearing in the top 10% of the MaxRel feature list vs. the total number of each feature type was also calculated. The results are illustrated in **Figure 2**. It is clear that all features of disorder status were in the top 80 features (10 / 10 = 1.000), followed by PSSM features (54 / 200 = 0.270), and SNT and PNT features (16 / 584 = 0.027). It can be concluded that the disorder status of the target protein sequences is more important than other properties.

As mentioned in Section III, 152 features were extracted as the optimal features. Among them, the numbers of SNT and PNT, PSSM and disorder status features were 115, 31, and 6, respectively. Similar with that of the top 80 features from the MaxRel feature list, the ratio of the 152 features in the mRMR feature list for each type was computed. The results are shown in **Figure 3**. These results showed that features of disorder status had the highest ratio, indicating that disorder status is also essential for classification. The detailed analysis of three feature types based on the above results can be seen below.

A. ANALYSIS OF DISORDER STATUS FEATURES

Among the three feature types (SNT and PNT, PSSM and disorder status), the relative ratio of disorder status was the highest, implying that sequence disorder is the principal influential factor for the CRISPR/Cas9 system. The top nine features in the MaxRel feature list and the top feature in the mRMR feature list are all disorder status features. It is true that the CRISPR/Cas9 gene editing progress does not require the participation of the proteins encoded by the target genes. However, this attribute of protein structure can be influenced by codon usage [22], suggesting the protein disorder property can somehow reflect certain DNA sequence features which we do not know so far. Our results demonstrated a strong correlation between this protein structure property with CRISPR efficacy, implying that some unknown DNA features can affect both the CRISPR efficacy and the protein structure. We also noticed that the ranks of disorder features in mRMR list were much lower than those in MaxRel list. This alteration may be attributed to the similarity of different disorder features, producing the corresponding redundancy which is considered when constructing the mRMR feature list.

B. ANALYSIS OF SNT AND PNT FEATURES

Apart from the disorder status features of the target sequence, another feature type, SNT and PNT, which describes the sequence characteristics of sgRNAs has also been screened out to be quite significant for the efficacy and accuracy of the CRISPR/Cas9 system. Based on mRMR method, such group of features turned out to be quite significant for CRISPR/Cas9 system (relative ratio 0.197 for optimal features). In the MaxRel feature list (before redundancy removal), two SNT and PNT features in this type followed the top nine disorder status features, validating the specific contribution of sgRNAs. As we all know, the core structure of sgRNA is a spacer sequence complementary to the targeted DNA sequence to guide the Cas9 or dCas9 proteins to genomic targets [60]. Various publications have reported that the sequence of the sgRNA may greatly affect the work efficiency of the CRISPR/Cas9 system [61]. By the MaxRel and mRMR feature lists, the 21st and 24th site of sgRNAs turn out to be quite significant for the CRISPR/Cas9 system. As we have mentioned above, there are two main basic principles for the sgRNAs to recognize and bind to the target sequence: initial 20 nt RNA sequence complementary pairing with the target sequence and the following three deoxyribonucleotides pairing with the conservative PAM (protospacer adjacent motif) district which usually contains specific nucleotide pattern: NGG [61]. Since the conservative PAM district usually turns out to be "NGG", the matching sgRNA sequence of such district should be "NCC", corresponding with the high rank of the feature that describes the frequency of "AC" in the 21st in sgRNAs. Furthermore, there are also publications reported the purine/pyrimidine composition near the 3' end of the spacer sequence (near the 20th site of sgRNAs) [62]. Another feature about the frequency of "G" in the 24th site of sgRNA may definitely affect the purine/pyrimidine composition near the 3' end of the spacer sequence, which may further influence the efficiency and accuracy of sgRNA and the whole CRISPR/Cas9 system. The relative ratios of this feature type for the top 80 features in the MaxRel feature list and optimal features (i.e., the top 152 features in the mRMR feature list) were quite different, validating the crucial regulatory role of such group of features for the CRISPR/Cas9 system.

C. ANALYSIS OF PSSM FEATURES

The PSSM features describe the evolutionary conservation of the target protein sequence, which may somehow reflect the conservation level of cDNA sequences in this study. The top PSSM feature in the mRMR feature list and top three PSSM features in the MaxRel feature list are all about the second amino acid site of the target sequence, meaning the 4th to 6th nucleotides in the target DNA fragments. It has been reported that the 14 nucleotides near (upstream of) the PAM region are crucial for the identification of the target protein in the CRISPR/Cas9 system [63]. However, our results found that the 5' end of the sgRNA-targeted DNA fragments is also important. In addition, it has been reported that the sequences of tracrRNAs show significantly high diversity while the sequences closely related to the CRISPR/Cas loci show high evolutionary conservation [64], emphasizing the association of the sequence conservation with CRISPR/Cas9 efficacy.

V. CONCLUSIONS

This study contributed a computational framework to predict the activity of sgRNAs binding to their target proteins. Based on the mRMR method, IFS method and support vector machine, we obtained a group of optimal features that may affect the activity of sgRNAs and influence the efficacy and accuracy of the CRISPR/Cas9 system. Based on the results, features describing the disorder status of the target proteins were significant for this system. Our newly proposed computational framework based on the support vector machine and mRMR method provides a new point of view for the famous CRISPR/Cas9 system, and the detailed analyses of the optimal features may help us gain insight into the underlying mechanism of this genome editing tool. Together, these results may further improve the efficacy and accuracy of the CRISPR/Cas9 system in a large number of practical applications.

ACKNOWLEDGMENT

(Lei Chen, ShaoPeng Wang, and Yu-Hang Zhang contributed equally to this work.)

REFERENCES

[1] C. Schlötterer, "Genes from scratch—The evolutionary fate of *de novo* genes," *Trends Genet.*, vol. 31, no. 4, pp. 215–219, Apr. 2015.

[2] H. P. Zhang and T. M. Yin, "Advances in lineage-specific genes," *Yi Chuan*, vol. 37, no. 6, pp. 544–553, Jun. 2015.

[3] M. J. Drake and P. Bates, "Application of gene editing technologies to HIV-1," *Current Opinion HIV AIDS*, vol. 10, no. 2, pp. 123–127, Mar. 2015.

[4] D. P. Weeks, M. H. Spalding, and B. Yang, "Use of designer nucleases for targeted gene and genome editing in plants," *Plant Biotechnol. J.*, vol. 14, no. 2, pp. 483–495, Feb. 2016.

[5] B. Petersen and H. Niemann, "Advances in genetic modification of farm animals using zinc-finger nucleases (ZFN)," *Chromosome Res.*, vol. 23, no. 1, pp. 7–15, Feb. 2015.

[6] Y.-I. Jo, H. Kim, and S. Ramakrishna, "Recent developments and clinical studies utilizing engineered zinc finger nuclease technology," *Cellular Molecular Life Sci.*, vol. 72, no. 20, pp. 3819–3830, Oct. 2015.

[7] B. Dupret and P. O. Angrand, "Targeted genome modifications using TALEN," *Med. Sci.*, vol. 30, no. 2, pp. 186–193, Feb. 2014.

[8] V. M. Bedell *et al.*, "*In vivo* genome editing using a high-efficiency TALEN system," *Nature*, vol. 491, pp. 114–118, Nov. 2012.

[9] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity," *Science*, vol. 337, no. 6096, pp. 816–821, Aug. 2012.

[10] P. D. Hsu, E. S. Lander, and F. Zhang, "Development and applications of CRISPR-Cas9 for genome engineering," *Cell*, vol. 157, no. 6, pp. 1262–1278, Jun. 2014.

[11] Y. Mei, Y. Wang, H. Chen, Z. S. Sun, and X.-D. Ju, "Recent progress in CRISPR/Cas9 technology," *J. Genet. Genomics*, vol. 43, no. 2, pp. 63–75, Feb. 2016.

[12] L. Qu, H. S. Li, Y. H. Jiang, and C. S. Dong, "The molecular mechanism of CRISPR/Cas9 system and its application in gene therapy of human diseases," *Yi Chuan*, vol. 37, no. 10, pp. 974–982, Oct. 2015.

[13] J. Lu *et al.*, "A redesigned CRISPR/Cas9 system for marker-free genome editing in *Plasmodium falciparum*," *Parasites Vectors*, vol. 9, p. 198, Apr. 2016.

[14] J. Peng, Y. Zhou, S. Zhu, and W. Wei, "High-throughput screens in mammalian cells using the CRISPR-Cas9 system," *FEBS J.*, vol. 282, no. 11, pp. 2089–2096, Jun. 2015.

[15] L. Xiao-Jie, X. Hui-Ying, K. Zun-Ping, C. Jin-Lian, and J. Li-Juan, "CRISPR-Cas9: A new and promising player in gene therapy," *J. Med. Genet.*, vol. 52, pp. 289–296, May 2015.

[16] P. I. Thakore *et al.*, "Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements," *Nature Methods*, vol. 12, pp. 1143–1149, Dec. 2015.

[17] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, "Genome engineering using the CRISPR-Cas9 system," *Nature Protocols*, vol. 8, pp. 2281–2308, Oct. 2013.

[18] T. Sakuma, A. Nishikawa, S. Kume, K. Chayama, and T. Yamamoto, "Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system," *Sci. Rep.*, vol. 4, p. 5400, Jun. 2014.

[19] B. P. Kleinstiver *et al.*, "Engineered CRISPR-Cas9 nucleases with altered PAM specificities," *Nature*, vol. 523, pp. 481–485, Jun. 2015.

[20] J. G. Doench *et al.*, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9," *Nature Biotechnol.*, vol. 34, pp. 184–191, Jan. 2016.

[21] M. Haeussler *et al.*, "Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR," *Genome Biol.*, vol. 17, p. 148, Jul. 2016.

[22] M. Zhou, T. Wang, J. Fu, G. Xiao, and Y. Liu, "Nonoptimal codon usage influences protein structure in intrinsically disordered regions," *Molecular Microbiol.*, vol. 97, no. 5, pp. 974–987, Sep. 2015.

[23] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[24] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, nos. 1–2, pp. 169–186, Sep. 2003.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[26] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Molecular Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[27] J. G. Doench *et al.*, "Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation," *Nature Biotechnol.*, vol. 32, pp. 1262–1267, Sep. 2014.

[28] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, Apr. 2012.

[29] L.-L. Hu *et al.*, "Prediction and analysis of protein palmitoylation sites," *Biochimie*, vol. 93, no. 3, pp. 489–496, Mar. 2011.

[30] Y. Zhou, N. Zhang, B.-Q. Li, T. Huang, Y.-D. Cai, and X.-Y. Kong, "A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis," *J. Biomolecular Struct. Dyn.*, vol. 33, no. 11, pp. 2479–2490, 2015.

[31] S. Niu *et al.*, "Predicting protein oxidation sites with feature selection and analysis approach," *J. Biomolecular Struct. Dyn.*, vol. 29, no. 6, pp. 650–658, 2012.

[32] Y. Cai, J. He, and L. Lu, "Predicting sumoylation site by feature selection method," *J. Biomolecular Struct. Dyn.*, vol. 28, no. 5, pp. 797–804, Apr. 2011.

[33] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucl. Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[34] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinform.*, vol. 7, p. 208, Apr. 2006.

[35] S. Niu, T. Huang, K. Feng, Y. Cai, and Y. Li, "Prediction of tyrosine sulfation with mRMR feature selection and analysis," *J. Proteome Res.*, vol. 9, no. 12, pp. 6490–6497, Dec. 2010.

[36] L. Chen, C. Chu, and K. Feng, "Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization," *Combinat. Chem. High Throughput Screening*, vol. 19, no. 2, pp. 136–143, 2016.

[37] L. Liu *et al.*, "Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection," *J. Biomolecular Struct. Dyn.*, vol. 35, no. 2, pp. 312–329, Feb. 2017.

[38] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, and Y.-D. Cai, "Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways," *Artif. Intell. Med.*, vol. 76, pp. 27–36, Feb. 2017.

[39] Q. Ni and L. Chen, "A feature and algorithm selection method for improving the prediction of protein structural class," *Combinat. Chem. High Throughput Screening*, vol. 20, no. 7, pp. 612–621, 2017.

[40] S. Wang, Y.-H. Zhang, J. Lu, W. Cui, J. Hu, and Y. D. Cai, "Analysis and identification of aptamer-compound interactions with a maximum relevance minimum redundancy and nearest neighbor algorithm," *Biomed. Res. Int.*, vol. 2016, Jan. 2016, Art. no. 8351204.

[41] J. Lu, S. P. Wang, Y.-D. Cai, and Q. Zhang, "Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm," *Current Bioinform.*, 2017, doi: 10.2174/1574893611666160608075753.

[42] B.-Q. Li, Y.-H. Zhang, M.-L. Jin, T. Huang, and Y. D. Cai, "Prediction of protein-peptide interactions with a nearest neighbor algorithm," *Current Bioinform.*, 2017, doi: 10.2174/1574893611666160711162006.

[43] Q. Zhang *et al.*, "Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm," *Combinat. Chem. High Throughput Screen*, vol. 20, no. 2, pp. 164–173, Dec. 2017.

[44] S. Wang, Y.-H. Zhang, G. Huang, L. Chen, and Y.-D. Cai, "Analysis and prediction of myristoylation sites using the mRMR method, the IFS method and an extreme learning machine algorithm," *Combinat. Chem. High Throughput Screening*, vol. 20, no. 2, pp. 96–106, 2017.

[45] L. Chen *et al.*, "Analysis of gene expression profiles in the human brain stem, cerebellum and cerebral cortex," *PLoS ONE*, vol. 11, no. 7, p. e0159395, 2016.

[46] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Multiple Classifier Systems*, vol. 3541, N. Oza, R. Polikar, J. Kittler, and F. Roli, Eds. Berlin, Germany: Springer, 2005, pp. 278–285.

[47] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Statist. Assoc.*, vol. 99, pp. 67–81, Jan. 2004.

[48] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," presented at the 5th Annu. Workshop Comput. Learn. Theory, Pittsburgh, PA, USA, Jul. 1992.

[49] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Res, Redmond, WA, USA, Tech. Rep. MSR-TR-98-14, 1998.

[50] I. H. Witten and E. Frank, Eds., *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2005.

[51] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta–Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[52] L. Chen, C. Chu, T. Huang, X. Kong, and Y.-D. Cai, "Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models," *Amino Acids*, vol. 47, no. 7, pp. 1485–1493, Jul. 2015.

[53] L. Chen, Y.-H. Zhang, M. Zheng, T. Huang, and Y.-D. Cai, "Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds," *Molecular Genet. Genomics*, vol. 291, no. 6, pp. 2065–2079, Dec. 2016.

[54] P.-W. Zhang, L. Chen, T. Huang, N. Zhang, X.-Y. Kong, and Y.-D. Cai, "Classifying ten types of major cancers based on reverse phase protein array profiles," *PLoS ONE*, vol. 10, no. 3, p. e0123147, 2015.

[55] N. Zhang *et al.*, "Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis," *PLoS ONE*, vol. 9, no. 9, p. e107464, 2014.

[56] L. Chen, B.-Q. Li, M.-Y. Zheng, J. Zhang, K.-Y. Feng, and Y.-D. Cai, "Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways," *Biomed. Res. Int.*, vol. 2013, Jul. 2013, Art. no. 723780.

[57] G. Huang *et al.*, "Exploring mouse protein function via multiple approaches," *PLoS ONE*, vol. 11, no. 11, p. e0166580, 2016.

[58] L. Chen, Y.-H. Zhang, T. Huang, and Y.-D. Cai, "Gene expression profiling gut microbiota in different races of humans," *Sci. Rep.*, vol. 6, Mar. 2016, Art. no. 23075.

[59] Y.-H. Zhang *et al.*, "Identification of the core regulators of the HLA I-peptide binding process," *Sci. Rep.*, vol. 7, p. 42768, Feb. 2017.

[60] P. Mali, K. M. Esvelt, and G. M. Church, "Cas9 as a versatile tool for engineering biology," *Nature Methods*, vol. 10, pp. 957–963, Sep. 2013.

[61] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9," *Nature*, vol. 507, pp. 62–67, Mar. 2014.

[62] T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander, "Genetic screens in human cells using the CRISPR-Cas9 system," *Science*, vol. 343, no. 6166, pp. 80–84, Jan. 2014.

[63] S. Lin, B. T. Staahl, R. K. Alla, and J. A. Doudna, "Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery," *Elife*, vol. 3, p. e04766, Dec. 2014.

[64] K. Chylinski, K. S. Makarova, E. Charpentier, and E. V. Koonin, "Classification and evolution of type II CRISPR-Cas systems," *Nucl. Acids Res.*, vol. 42, no. 10, pp. 6091–6105, 2014.

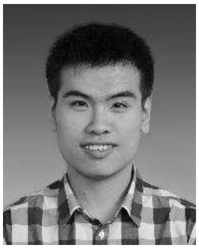
LEI CHEN received the B.S. degree, the M.S. degree in operational researches, and the Ph.D. degree in system analysis and integration from the East China Normal University, in 2004, 2007, and 2010, respectively.

Since 2010, he has been with the College of Information Engineering, Shanghai Maritime University, as an Associate Professor. His interests include bioinformatics, computational biology, graph theory, and algorithm design.

Dr. Chen is the member of the China Computer Federation and the Chinese Association for Artificial Intelligence. He is the Editorial Board member of Current Bioinformatics.

SHAOPENG WANG was born in Lanzhou, China, in 1989. He received the B.S. and M.S. degrees in chemistry from the Northwest Normal University in 2012 and Lanzhou University in 2015. He is currently pursuing the Ph.D. degree in bioinformatics and system biology with the School of Life Sciences Shanghai University, China.

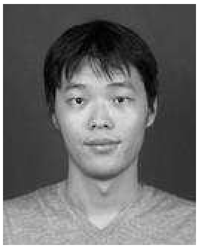
From 2013 to 2017, his research interest includes the predicting post-translational modification sites from protein sequences by using machine learning algorithms, mining novel disease-related genes based on protein-protein network methods, and classifying functional macromolecules, such as RNAs and proteins and drug-like molecules.



YU-HANG ZHANG was born in Jinzhou, China, in 1992. He received the B.S. degrees from the Medical Laboratory, Shanghai Jiaotong University Medical School, in 2014. He is currently pursuing the Ph.D. degree with the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. He has authored over 20 articles. His research interests include machine learning, liquid biopsy, and tumor immunotherapy. He was a recipient of Merit Student at the University of Chinese Academy of Sciences in 2017.

JIARUI LI was born in China in 1982. He received the B.S. degree in biology and the Ph.D. degree in genetics from Fudan University, Shanghai, China, in 2005 and 2012, respectively.

From 2012 to 2016, he was a Post-Doctoral Research Fellow with the Dr. J. Chen's Team, Simon Fraser University, Burnaby, BC, Canada. Since 2009, he has been a Lecturer with the School of Life Science, Shanghai University, Shanghai. His research interests include comparative genomics through applying bioinformatics tools and cutting-edge sequencing technologies.



ZHI-HAO XING was born in Jining, China, in 1990. He received the Ph.D. degree in genetics from the Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.

His research interest includes the fundamental study of genetics and epigenetics of human diseases. His awards and honors include the National Scholarship for Graduate Students and School Miyoshi Students.

JIALIANG YANG received the Ph.D. degree from the Department of Mathematics, National University of Singapore, in 2009. From 2010 to 2011, he was an Assistant Professor with the CAS-MPG Partner Institute for Computational Biology, China. He moved to USA and became a Post-Doctoral Fellow at the Department of Basic Sciences, Mississippi State University. Since 2013, he has been a Post-Doctoral Fellow at the Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai. He has published over 40 peer-reviewed articles. His main research areas include bioinformatics, machine learning, aging, and evolutionary biology.

TAO HUANG received the B.S. degree in bioinformatics from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in bioinformatics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, in 2012. Since 2014, he has been an Associate Professor and the Director of Bioinformatics Core Facility with the Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

From 2012 to 2014, he was a Post-Doctoral Fellow with the Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York City, USA. His research interest includes bioinformatics, computational biology, systems genetics, and big data research.

He has published over 100 articles. His researches have been cited for over 3000 times with an h-index of 26 and an i10-index of 64. He has been a reviewer for over 20 journals and an editor/guest editor for seven journals and books.

YU-DONG CAI has been a Professor in bioinformatics with the School of Life science, Shanghai University since 2015. His main interests cover various areas of systems biology and bioinformatics, such as protein–protein interaction, disease biomarkers prediction, drug-target interaction, and protein functional sites prediction. He has published over 200 peer-reviewed scientific papers, including invited reviews. His researches have been cited for over 7500 times, with an h-index of 51. He is the Editorial Board Member of *Biochimica et Biophisica Acta-Proteins and Proteomics*, *Biochemistry Research International*. He has been a Guest Editor for *Computational Proteomics, Systems Biology and Clinical Implications*, *Biochimica et Biophysica Acta-Proteins and Proteomics*.