A Computational Method for Identification of Functional SNPs in Human Noncoding Genome Regions Based on Multi-feature Mining

*Rong Li, Zhi-e Lou**

Telecommunication and Networks National Laboratory, Nanjing University of Posts and Telecommunications, China
lirong@njupt.edu.cn, louzhie@njupt.edu.cn

Abstract

Single Nucleotide Polymorphism (SNP) is the variant on a single nucleotide in the genome. Functional SNP, as one of the most important molecular markers in disease research, has been widely used in various research fields, such as tumor pathogenesis, disease diagnosis and treatment, prognostic evaluation, drug development, etc. The number of functional SNPs in noncoding genome regions is much more than that in coding regions, and their detection is more difficult. In this work, a multi-feature mining based computational method is proposed to predict the functional SNPs in human noncoding genomes. We first analyzed the sequence properties, evolutionary conservation properties and epigenetic modification signal properties of the sample SNPs. Statistical methods together with multiple annotation data from genomes and epigenetics were used to mine high-dimensional discriminative features subsequently. In particular, the allele-specific features were designed to distinguish the function of SNPs with close locations. The random forest method was used to conduct feature dimension reduction and classification. The 10-fold cross-validation result showed the Area Under the Receiver Operating Characteristic Curve (AUC) of our method improved by 16.9% and 43.4% over existing methods GWAVA and CADD, respectively, illustrating that the allele-specific based features can help to distinguish functional and netural SNPs with near locations.

Keywords: Feature mining, Evolutionary conservation, Epigenetic modification, Feature selection

1 Introduction

Though the human populations vary widely in phenotypes, differences in genomes are very small and they are called genetic variants. The presence of genetic variants causes differences of the human population and individuals in body constitution, athletic ability, disease susceptibility, intelligence, and psychology. Single Nucleotide Polymorphism (SNP) is one of the most common types of genetic variants. The SNP has a high density and stability, and it's easy to be genotyped, automated, and launched for large-scale analysis [1]. Therefore, it has become the third-generation genetic marker and is widely used in research areas such as population genetics, pharmaceutical industry, forensic medicine, tumor pathogenesis, and diagnosis and treatment of complex genetic diseases [2].

The SNP distributes evenly across the whole genome. It's estimated that there are about 10 million SNPs in the human genome, with about one in every 290 bases. Among the massive SNPs, only a few can affect gene expression and they are called functional SNPs. When performing large-scale association analyses of complex diseases, the use of reliable functional SNPs instead of genome-wide SNPs as genetic markers can reduce the workload of sequencing and analysis, thus accelerating the study of pathogenic gene localization. Functional SNPs located in coding regions can effectively change the amino acids, thus affecting the structure and function of the encoded protein, so their detection is much easier. However, studies have shown that more functional SNPs are located in non-coding genome regions [3], which can affect gene expression by influencing gene splicing, transcription factors (TFs) bindings, mRNA degradation, and non-coding RNA sequences [4-5]. Because functional SNPs in non-coding regions don't alter the amino acids of encoded proteins, and the way they affect gene expression has still not been fully understood, their detection is more difficult.

It's very necessary to establish prediction models of non-coding functional SNPs via computer-aided approaches. The establishment of computational models can help biological researchers to quickly access candidate SNPs from massive background SNPs for further analysis, thus speeding up the search for pathogenic genes. Current prediction and analysis methods of non-coding functional SNPs can be roughly divided into two categories: functional elements based methods and machine learning based methods.

Functional elements based methods mainly rely on functional elements data annotated by experiments, and they judge the functionality of an SNP by analyzing its association (mainly the location relationship) with those functional elements. GenomeRunner [6] determines the function of SNPs based on the position relationships between SNPs and the enriched regions of epigenomic features. HaploReg [7] annotates the effect of SNPs on gene modules based on the ChIP-Seq data of TFs.

Machine learning based methods collect functional and control SNP samples and then train classifiers to predict their functions. DANN [8] compares variants that survive natural selection with the simulated mutations, and then uses the deep learning method to predict the function of SNPs by

*Corresponding Author: Zhi-e Lou; E-mail: louzhie@njupt.edu.cn
DOI: 10.53106/160792642023052403021

combining multiple genome-related annotation data. GWAVA (Genome-Wide Annotation of Variants) [9] uses a range of variant-related annotation data of different types and at different genome scales to investigate whether the effective combination of functional elements, gene backgrounds, and genome-wide properties can be used to identify possible functional variants. CADD (Combined Annotation–Dependent Depletion) [10] predicts the pathogenicity of genetic variants in human genome by integrating multiple annotation data from the Ensembl Variant Effect Predictor (VEP) [11], ENCODE program [12], and UCSC Genome Browser tracks [13] into one metric - C-score, which has been normalized to the range 1-99, and CADD has provided pre-computed C-scores for 860 million human single nucleotide variants. CERENKOV3 [14] introduces the clustering features and molecular network features of the SNP clusters to predict functional SNPs by using the gradient boost framework based decision tree. Ramsey et al. use a naive Bayes-like framework to combining three quantities for SNPs, thus prioritizing candidate noncoding functional SNPs [15].

This paper addresses the prediction problem of functional SNPs in noncoding regions of the human genome. We proposed a sequence recognition optimization algorithm for functional SNPs based on the sequence statistical and biological properties of sample SNPs, and then built a prediction model based on multiple features fusion to discriminate functional SNPs from the neutral ones. The most important part of our method was the establishment of feature engineering, and numerous features of different categories were extracted, including evolutionary conservation scores, histone modification signals, DNase signals, DNA conformational features, etc. In particular, the allele-specific features were designed to distinguish the function of the SNPs with close locations, and it was the existence of these features made our method perform better on data that had been collected through the most strict method thought by Ritchie et al. [9]. We think that our work could provide theoretical support for further understanding of the pathogenic mechanism of functional variants. And we also hope such a method could help geneticists to rapidly assess likely functional SNPs from a flood of genetic polymorphisms for further studies in pathogenic gene mapping and drug development.

2 Materials and Methods

2.1 Data Collection

Functional SNPs were obtained from the Genome-Wide Association Study (GWAS) Catalog database [16]. GWAS Catalog is a publicly available database of fine published GWAS findings by artificial selection. The disease/trait-associated SNPs in GWAS Catalog refer to the high association SNPs with GWAS *p*-values less than 10^{-5} . All the 24,263 disease/trait-associated SNPs contained in GWAS Catalog (up to July, 2019) were downloaded first. Then we mapped all these functional SNPs to dbSNP database (dbSNP 146, the human genome build 38) to ensure the data validity, and then SNPs located in noncoding regions were picked. Finally, 17,700 SNPs were left as the positive samples, and most of them located within the intronic and intergenic regions (see Figure 1).

For the control set, following the most stringent method of constructing negative samples described by Ritchie et al. [9], all common SNPs, that were SNPs with minor allele frequency (MAF) greater than 0.5, within 2kb upstream and downstream of the collected functional SNPs were downloaded. Then, according to the quality control method we proposed which was shown in Figure 2, all potential functional SNPs were removed through the 5-steps "remove" operation. Finally, the remaining 157,057 neutral SNPs constituted our negative dataset.

2.2 Feature Extraction

2.2.1 Sequence Context based Features

To examine the differences in nucleotide compositions between functional and neutral SNP sequences, we counted the position-specific nucleotide distributions of SNP sequences from 20bp upstream to 20bp downstream of the SNP sites. As the results shown in Figure 3, the distribution content of bases A and T at both upstream and downstream positions of functional SNP sequences are higher than that of the neutral ones, while bases C and G perform the opposite situation. The commonly used statistical methods for extracting biological sequence properties, including GC content [19], *k*-mers position weight matrix (PWM) scores [20], and *k*-mers statistical discrete increment [21], were used to characterize the differences of nucleotides between positive and negative samples.

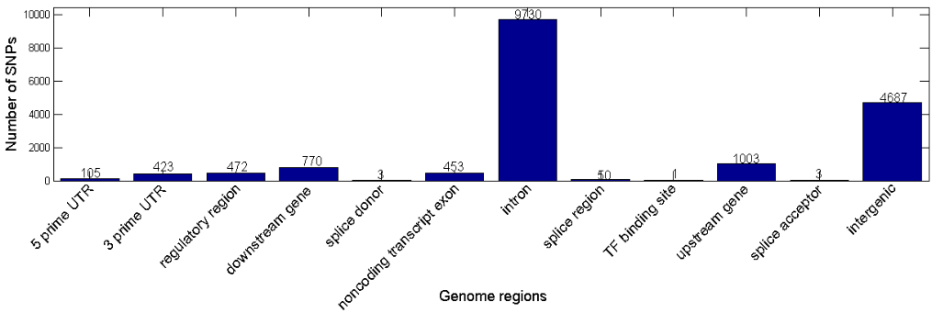


Figure 1. The distribution of genomic locations of noncoding functional SNPs in GWAS Catalog

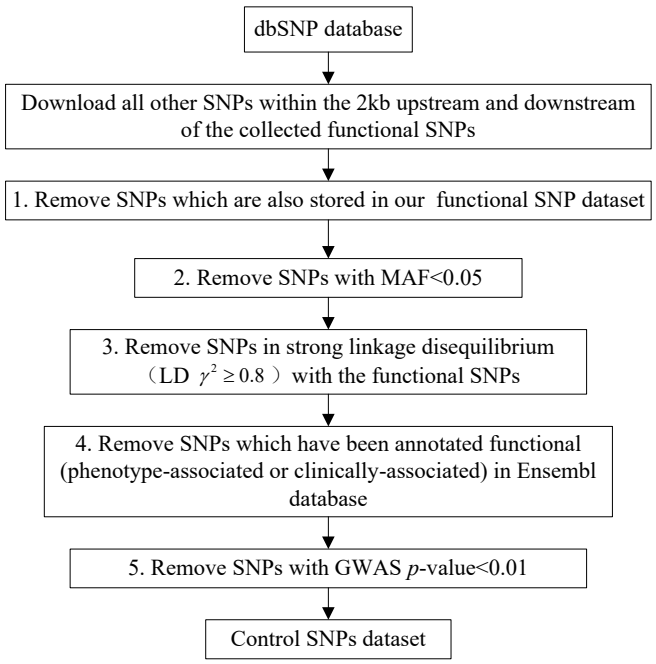


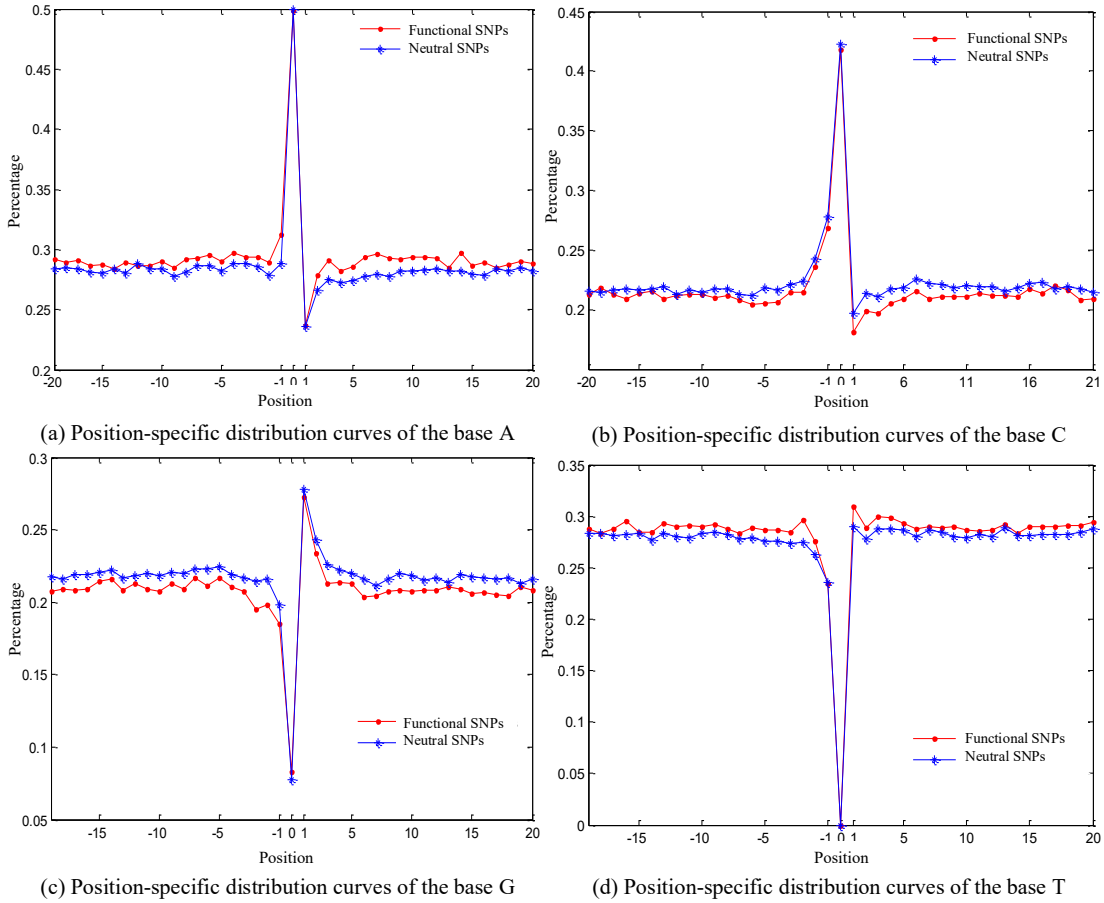
Figure 2. The collection steps of the control set. Linkage disequilibrium data were from HapMap (HapMap rel27, phases I+II+II) [17], GWAS association *p*-values were from GWAS central [18]

2.2.2 Biological Properties based Features

Biological properties based features were divided into two main categories: position based features and allele-specific based features. Position based features investigate the properties of the genome regions where SNPs locate, while allele-specific based features study the properties of impacts caused by SNPs.

Position based features

Evolutionary conservation. Conserved regions in sequences are generally considered as functional sites, because noncoding functional elements have higher conservation than their surrounding nonfunctional sequences [22]. Therefore, the evolutionary conservation score of the SNP location can be considered as a feature to discriminate the functional SNP. The phastCons and phyloP scores [23] together with the GERP [24] score were used to characterize the conservative features. The phastCons and phyloP scores at the single nucleotide resolution in the UCSC database [13] are given in BigWig [25] format. Using the format transformation function "bigWigAverageOverBed" provided in UCSC, the BigWig format phastCons and phyloP values can be converted to different conservation metrics of sequences at any length, including the maximum, minimum, mean, and total conservation scores. In our work, the total phastCons and phyloP scores of sequences with length of 41bp (20bp upstream and downstream of SNP sites) were used to represent the degree of conservation of the regions where SNPs locate.



(a) Position-specific distribution curves of the base A

(b) Position-specific distribution curves of the base C

(c) Position-specific distribution curves of the base G

(d) Position-specific distribution curves of the base T

Figure 3. Position-specific distribution profiles of nucleotides for both functional and neutral SNPs, X-axis represents the nucleotide position with 0 being the SNP site

The GERP scores were retrieved from GERP++ software [26]. GERP++ provides GERP scores corresponding to each nucleotide position on the 24 chromosomes of the human genome build 19 (hg19). There are two scores per nucleotide position in GERP++, namely, the neutral evolutionary rate (gerpN) and the rejected substitution rate (gerpS). The gerpN and gerpS scores of all collected SNPs were got by mapping their positions to hg19. The Mann–Whitney U test showed that the two scores of functional and neutral SNP show significant statistical differences (gerpN: p=0.0, gerpS: p=6.72×10-4), so these two scores also served as the discriminative features of our prediction method.

Epigenetic modification features. Epigenetic modifications include DNA methylation and histone modifications. Methylation and histone modifications are found to play regulatory roles by affecting the affinities of TFs and the promoters of structural genes [27]. Specific gene regions, such as enhancers, promoters, as well as some genetic antibodies, have different histone modification signals [28]. Thus, the histone modification signal has been widely used in the prediction of cis-acting elements such as enhancers and TFs.

The three important methylation and histone modification signals, namely H3k4me1, H3k4me3, and H3k27ac, for six human cell lines (Gm12878, Nh1f, Hsmm, Huvec, K562, and Nhek) were downloaded from UCSC database. The format and computing method of the three modification signals were the same as phastCons and phyloP scores. We selected the maximum and total expression levels of histone modification singals in a certain range of length to characterize the difference between functional and neutral SNPs, and here both the maximum and total expression levels were the corresponding maximum of the six cell lines. Figure 4 shows the statistical differences in modification signals between functional and neutral SNPs using the Mann-Whitney U test. We can see that the expression levels of H3k4me1, H3k4me3, and H3k27ac perform significant differences (1.0×10-62<p<1.0×10-50) within the survey regions. Because all three modification signals showed significant differences within the survey regions, we chose a window with length 41bp to calculate the histone modification feature to keep concordance with the choice we made in computing evolutionary conservations. Finally, the maximum and total expression levels of the three histones in a region of length 41bp (SNP site+/- 20bp) were extracted as features to identify functional SNPs.

DNase singal. Deoxyribonuclease I (DNaseI) is an endonuclease that can digest DNA and produce mono-or oligodeoxynucleotides. Studies have found that functional DNA fragments or protein binding regions on chromatin are overlapped with DNase hypersensitive sites (DHSs) [29]. Therefore, the localizations of DHSs have become effective means to accurately identify regulatory elements on chromatin and determine the binding regions of regulatory proteins [30]. The UCSC database also provides DNase signal profiles, and the maximum and total DNase expression levels within a certain region can be obtained by the same method of computing conservation scores.

The Mann-Whitney U test was used to select the appropriate statistical length for the DNase signal. Figure 5 shows that the significant difference in total DNase expression level increases significantly with sequence length, while the maximum DNase expression level decreases, but both of them show very good statistical significance (p<1.0×10-170). The same as above, we selected the maximum and total expression levels of DNase within a region of length 41bp (SNP site+/- 20bp) as discriminative features to distinguish functional SNPs.

Another position based feature is regulatory elements hit, which is an index of enrichment levels of functional elements around the SNP site. In our previous work, we found this feature is very effective for predicting regulatory SNPs [31]. So, the regulatory elements hit was also used here as the discriminative feature, and it can be computed following our previous work [31].

Allele-specific based features

Allele-specific based features mainly examine the influences of SNP variants on DNA structure and energy, etc.

The DNAshape structure properties. DNAshape [32] is an online prediction tool for four structural features of DNA, including minor groove width (MGW), roll, propeller twist (ProT), and helix twist (HelT). The four structural properties of DNA fragments with arbitrary length can be calculated from the online DNAshape tool (https://rohslab.usc.edu/DNAshape/). Totally, 5bp nucleotide positions were under the influence of the SNP variant, that is 2 bp upstream and 2 bp downstream of the SNP site.

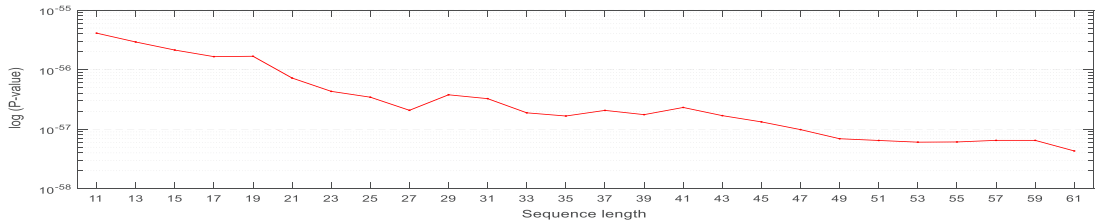
The total effect of an SNP variant on sequence structure was adopted as the discriminative feature, and it's measured by the Euclidean distance [33-34] between the two vectors as follows:

$$\Delta Ds_{sum} = \sqrt{\left(\overrightarrow{Ds}_r - \overrightarrow{Ds}_a\right)\left(\overrightarrow{Ds}_r - \overrightarrow{Ds}_a\right)^T}. \quad \textbf{(1)}$$

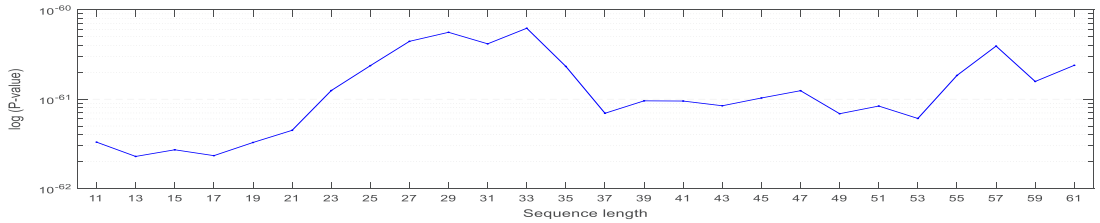
where $\overrightarrow{Ds_r}$ is the nucleotide structure feature vector of the reference sequence with length 5 got by DNAshape, and $\overrightarrow{Ds_a}$ is the structure feature vector of the alternative sequence, the symbol T represent the transpose of a vector.

In addition to the abovementioned DNAshape structure properties, the two types of features we previously used in regulatory SNPs detection [20], that are differences in conformational and thermodynamic properties and differences in hydroxyl radical cleavage patterns, were also used here. The methods of computing these two feature categories can be found in our previous publication [20]. In this work, the original property values of DNA conformation based on dinucleotides were from Goni et al. [35], and property values of DNA thermodynamics were from SantaLucia et al. [36].

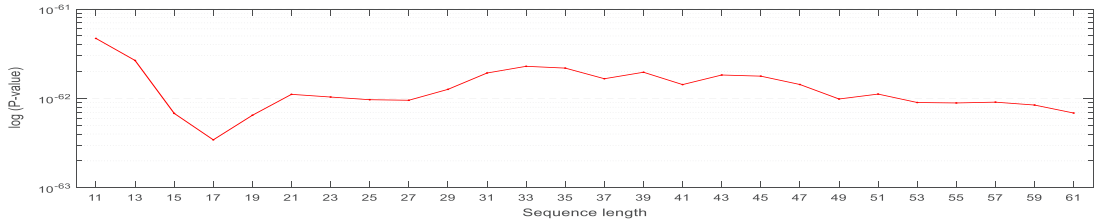
To sum up, the final feature set we extracted was shown in Table 1.



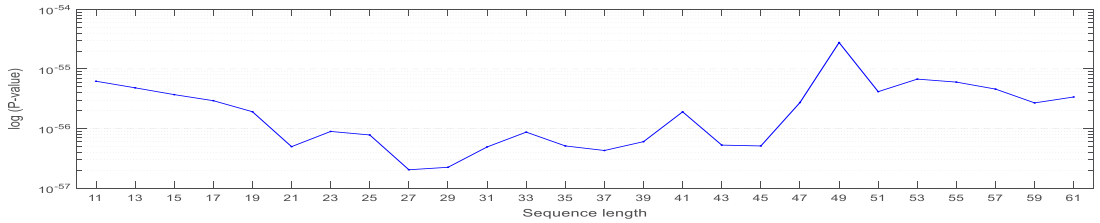
(a) Statistical difference of the total H3k4me1 expression levels



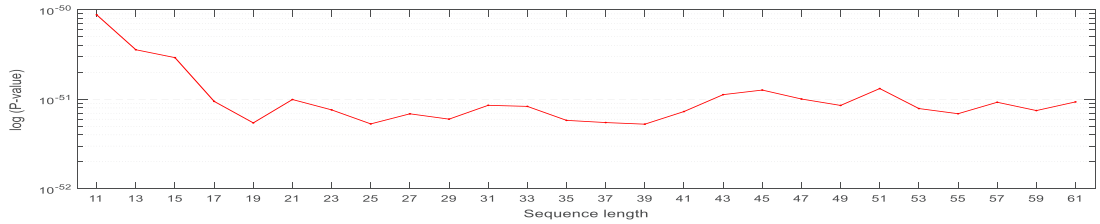
(b) Statistical difference of the maximum H3k4me1 expression levels



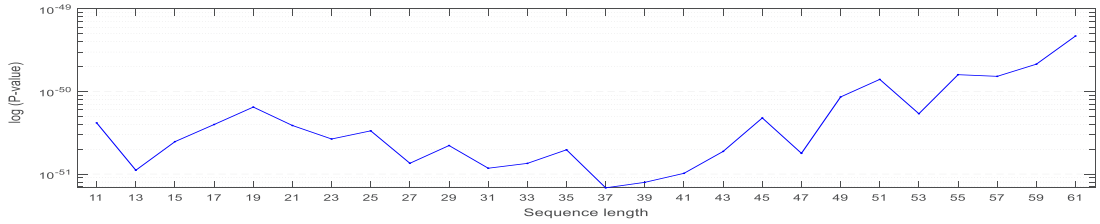
(c) Statistical difference of the total H3k4me3 expression levels



(d) Statistical difference of the maximum H3k4me3 expression levels

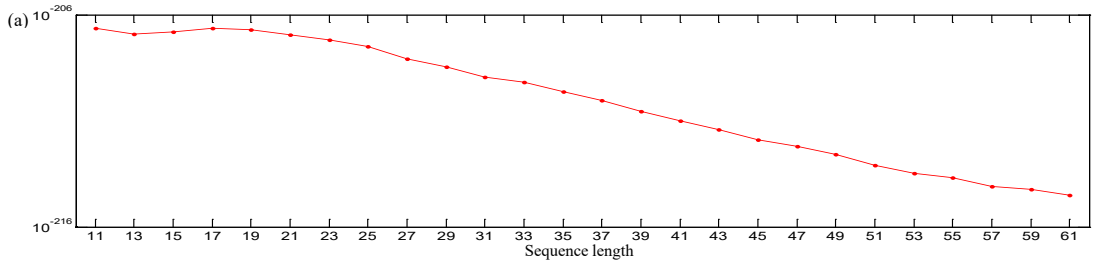


(e) Statistical difference of the total H3k27ac expression levels

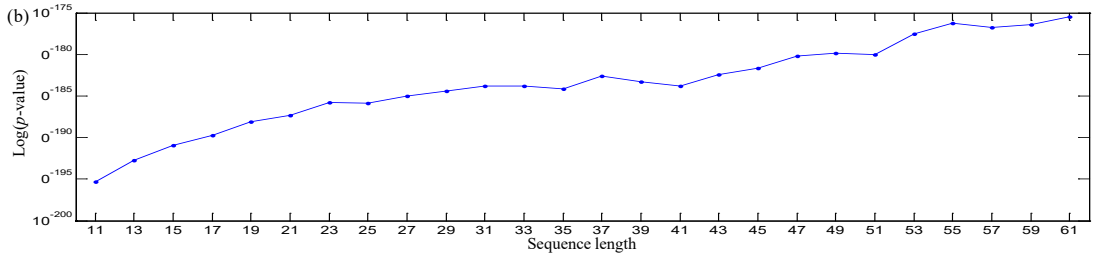


(f) Statistical difference of the maximum H3k27ac expression levels

Figure 4. Statistical difference of epigenetic modification signals between functional and neutral SNPs, X-axis represents the sequence length centered on the SNP site



(a) Statistical difference of the maximum DNase expression levels



(b) Statistical difference of the total DNase expression levels

Figure 5. Statistical difference of DNase expression levels between functional and neutral SNPs, X-axis represents the sequence length centered on the SNP site

2.3 Feature Selection and Classification

Random forest is a statistics-based machine learning method proposed by Breiman in 2001 on the base of decision tree. Random forest gets forest by constructing a certain number of mutually independent CART decision trees. So, when a new sample comes, each decision tree in the forest will predict the sample label, and the final classification result is determined according to the voting rule.

In addition to classification, random forest can also be used to prioritize features. The Bagging [37] algorithm is used in random forest when sampling. The main idea of Bagging is that only part of the training samples is selected using the put-back sampling method for training each decision tree. Therefore, there are some samples in the initial training set that can never be drawn, and they are called Out-Of-Bag (OOB) data. The OOB data can be used to evaluate the generalization error of the performance of random forest and to prioritize features. The OOB data prioritizes features by adding random noise to a feature, and the greater the OOB error is, the higher the feature importance will be. Thus, the feature causes a greater influence on classification result.

Assuming that Ntree subsets of the original sample set are obtained through bootstrap sampling, in other words, Ntree decision trees are established. Then the importance measurement (denoted as D_j) of feature x_j ($1 \le j \le K$, K is the total number of features) can be calculated as follows:

(1) For No. b=1 subset, denote its OOB data as L_b^{oob};

(2) Build a CART decision tree T_b to classify L_b^{oob}, and record the number of samples that have been correctly classified as R_b^{oob};

(3) For feature x_j ($1 \le j \le K$), do

a) Assign random noise to all components of x_j in L_b^{oob}, and denote it as L_{bj}^{oob};

b) Use T_b to classify L_{bj}^{oob}, and record the number of samples that have been correctly classified as R_{bj}^{oob};

(4) For b = 2, ..., *Ntree*, repeat steps (1) to (3);

(5) Calculate D_j: $D_j = 1/Ntree \sum_b \left(R_b^{oob} - R_{bj}^{oob}\right)$.

2.4 Performance Evaluation

Sensitivity (Sn) and specificity (Sp) are commonly used evaluation indexes for unbalanced data classification problems. Sensitivity is the identification rate of the model on positive samples, while specificity is the classification accuracy rate of the model on negative samples. In addition, the Receiver Operating Characteristic curve (ROC curve), a comprehensive indicator of the continuous changes in sensitivity and specificity, is also used to evaluate the performance of our algorithm. The Area Under the ROC Curve (AUC) is a quantitative evaluation of the ROC curve, and the larger the AUC value is, the better the classification performance will be.

3 Result

3.1 Performance and Analysis

The dataset we established was unbalanced with many more neutral SNPs than functional SNPs. The random split based ensemble learning method proposed by Sun et al. [38] was used to deal with the unbalanced data classification problem. Then, for each balanced subset, random forest was adopted to classify functional SNPs. The Dempster-Shafer (DS) evidence theory [39] was used to fuse the results of decision layer to obtain the final classification results subsequently. When using the random split based ensemble learning method to deal with unbalanced data, the samples of the majority class in training dataset were randomly split

into 9 (157,057/17,700=8.87) portions, and then each portion was combined with functional SNPs data in training dataset. So, there are nine balanced training subsets, and nine random forest classifiers were needed. The random forest classifier is not very sensitive to the model parameters, so we set the number of trees in random forest at *ntree*=500, and the number of features selected for each decision tree was set at $mtry = \lfloor \sqrt{D} \rfloor$ (D is the feature dimension). In order to obtain more objective results, 10-fold cross-validation was used, and to avoid the deviation error caused by the fixed data split pattern in data balancing, the 10-fold cross-validation procedures were repeated five times to average the results. We selected the sequence with length of 201bp that centered on SNP site (SNP site +/- 100bp) to build the prediction model. The 2-mers PWM scores were calculated by sliding windows of length 10 and overlap 5. So, the number of PWM score features was 4 × ((100 − 10) / (10 − 5) + 1) = 76. The total dimension of the feature vector corresponding to Table 1 is 106.

The random forest dimensional reduction method was used to investigate the prediction accuracy of the model under different feature dimensions (from 1 to 106), and the result was shown in Figure 6. In summary, when the feature dimension is less than 20, the overall recognition accuracy (corresponding to the evaluation index AUC) increases with the feature dimension increases. While after the feature dimension greater than 20, the overall recognition accuracy fluctuates slightly and evenly. The specificity and sensitivity of our method fluctuate widely with the feature dimension increases. When the feature dimension is 85, the method gets the largest AUC with 0.762, and the highest sensitivity with 71.32%, and the corresponding specificity is 72.34% at this time.

3.2 Comparisons Against Other Methods

We compared the prediction effects of our proposed model with GWAVA and CADD. The web server of GWAVA has provided pre-calculated scores from 0 to 1 for all single nucleotide variants in the Ensembl database [40] (release 70), and a higher score indicates a greaterlikelihood of the variant being functional. For each SNP, GWAVA provides three different scores corresponding to three different control sets, which are denoted as “unmatched score”, “TSS score” and “region score”, while CADD provides the normalized C-score.

Table 1. The final feature set we extracted for building the prediction model

Feature category		Feature name	Description	Feature dimension
Sequence context based feature		GC content	Contents of bases G and C in DNA sequences	1
		k-mers PWM	2-mers PWM scores calculated using the sliding window with length 10 and overlap 5 for weights from both positive and negative samples	4×(*L*/5-1) [a]
		k-mers discrete increment	Discrete increment values calculated by 1-mers and 2-mers.	2
Biological properties based feature	Position based feature	Regulatory elements hits	Enrichment levels of functional elements around the SNP site.	1
		Evolutionary conservation score	Conservation score calculated from phastCons, phyloP and GERP++.	4
		Epigenetic modification features	The maximum and total expression levels of methylation and histone modification singal within the region(SNP site+/- 20bp).	8
	Allele-specific based feature	DNAshape	Changes caused by the SNP variant on 4 DNA structures.	4
		DNA conformational feature	Changes caused by the SNP variant on DNA structural properties.	6
		DNA thermodynamic features	Changes caused by the SNP variant on DNA energy properties.	3
		Hydroxyl radical cleavage pattern	Changes caused by the SNP variant on hydroxyl radical cleavage patterns.	1

[a] *L* here is the length of flanking sequences around the SNP site.

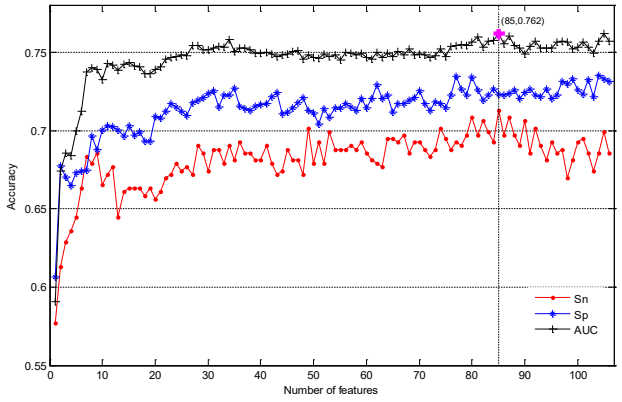


Figure 6. The performance of our proposed method

We obtained all pre-calculated scores for the 17,700 functional SNPs and 157,057 neutral SNPs in our dataset from the web servers of GWAVA and CADD. The three scores of GWAVA are the probabilities of random forest classifiers predicting an SNP being functional, and scores range from 0 to 1. So, if a threshold is given, GWAVA can determine directly whether an SNP is functional or not. While CADD gives C-scores from 1-99 that are highly associated with various types of single nucleotide variants or small insertion/deletion variants. These C-scores can't be directly used to determine the potential function of an SNP, but rather discriminate between pathogenic and benign variants by training support vector machine (SVM) with linear kernel. We used the same method as that used in CADD publication to handle C-scores, meaning linear kernel SVM will be built to convert C-scores to prediction results. As the same with our method, the result of 10-fold cross-validation is considered as the final prediction result for CADD.

The final comparison results of GWAVA, CADD and our proposed method are shown in Figure 7, where the prediction result of our method is the result of random forest with the feature dimension being 85. For GWAVA, the «region score» has the best prediction accuracy and the corresponding AUC is 0.65. CADD achieves prediction accuracy of AUC of 0.53. While the AUC of our proposed method is 0.76, improved by 16.9% and 43.4% over GWAVA and CADD respectively. Considering reasons for this phenomenon, GWAVA and CADD methods obtain numerous features from genome annotation data, ENCODE program, and many other high-throughput sequencing data, but these features are mostly positioning attributes, such as DNase-seq peak data, RNA polymerase binding profiles, TFs binding profiles etc. These features are more effective for finding SNPs associated with genomic functional elements or SNPs located within genome functional regions. However, SNPs located within the functional regions do not necessarily exhibit function due to the degenerative effects of the genes. For this reason, the features GWAVA and CADD used are not so effective to discriminate between functional and neutral SNPs that both located within functional genome regions. The negative samples we used to establish the dataset are SNP samples picked from 2kb upstream and downstream of the collected functional SNPs sites. So, our negative dataset will unavoidably contain a lot of SNPs that locate within genome functional elements but have not been found to have any trait or disease associations yet, and the presence of these SNPs makes GWAVA and CADD with poor prediction power for our dataset.

In our method, sequence context based features and position based features are drawn to find the characteristics of the SNPs location, which has the same thought as CADD and GWAVA. In addition, the allele specific based features are designed to distinguish functional SNPs from near neutral SNPs, thus making our method more effective than GWAVA and CADD.

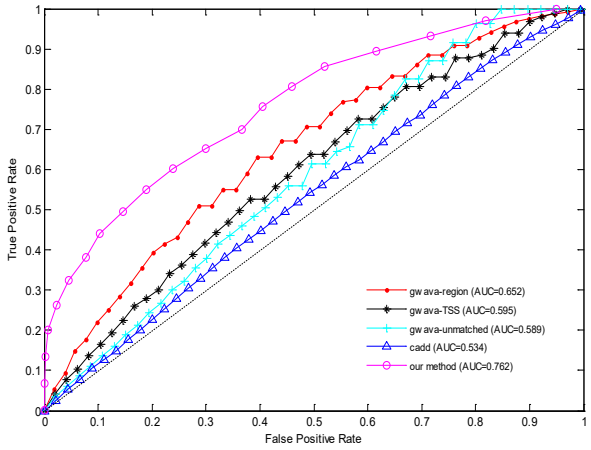


Figure 7. The ROC curves of GWAVA, CADD and our method

4 Conclusion

The number of functional SNPs in noncoding regions of the human genome is much more than that in coding regions. And their detection is more difficult, as the functions of most of the genome noncoding regions haven't been fully understood. Therefore, computational methods with rapid and large-scale prediction capabilities are in urgent demand. With the help of many publicly available databases and tools, lots of biological or statistical features associated with SNPs function are mined, and on the basis of these features, a prediction model for identifying functional SNPs is built. The result shows that our prediction method is more effective than other methods on a very strict dataset.

Considering that the occurrence and development of diseases involve complex regulatory mechanisms at multiple levels and multiple omics, such as genomic, transcriptome, proteome, metabolome, etc. So, the further work is to introduce more omics data into the establishment of feature engineering. We think the joint analysis of multi-level and multi-omics data can contribute to a more systematic and comprehensive understanding of the biological behavior of phenotypes and diseases, and further provide new clues for finding valuable disease markers and exploring disease-related mechanisms.

Acknowledgement

We are grateful to our colleagues in Telecommunication and Networks National Laboratory, Nanjing University of

Posts and Telecommunications, for their help during the course of this work, in particular Prof. Xiao-yong Yan for his critical suggestions.

Funding. This research was sponsored by grants from the Scientific Research Foundation of Nanjing University of Posts and Telecommunications (NUPTSF Grant No. NY218143 and No. NY218141), the National Fund incubation project of Nanjing University of Posts and Telecommunications (NUPTSF Grant No. NY219116) and the Natural Science Fund for Colleges and Universities in Jiangsu province (No. 22KJB520024).

References

[1] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. P. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, E. S. Lander, Large-scale Identification, Mapping, and Genotyping of Single-nucleotide Polymorphisms in the Human Genome, *Science*, Vol. 280, No. 5366, pp. 1077-1082, May, 1998.

[2] T. Fadason, S. Farrow, S. Gokuladhas, E. Golovina, D. Nyaga, J. M. O'Sullivan, W. Schierding, Assigning Function to SNPs: Considerations When Interpreting Genetic Variation, *Seminars in Cell & Developmental Biology*, Vol. 121, pp. 135-142, January, 2022.

[3] T. Lappalainen, E. T. Dermitzakis, Evolutionary History of Regulatory Variation in Human Populations, *Human Molecular Genetics*, Vol. 19, No. R2, pp. 197-203, October, 2010.

[4] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M.-Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, M. Snyder, Variation in Transcription Factor Binding Among Humans, *Science*, Vol. 328, No. 5975, pp. 232-235, April, 2010.

[5] J. Kroliczewski, A. Sobolewska, D. Lejnowski, J. F. Collawn, R. Bartoszewski, MicroRNA Single Polynucleotide Polymorphism Influences on MicroRNA Biogenesis and mRNA Target Specificity, *Gene*, Vol. 640, pp. 66-72, January, 2018.

[6] M. G. Dozmorov, L. R. Cara, C. B. Giles, J. D. Wren, GenomeRunner Web Server: Regulatory Similarity and Differences Define the Functional Impact of SNP Sets, *Bioinformatics*, Vol. 32, No. 15, pp. 2256-2263, August, 2016.

[7] P. Kheradpour, M. Kellis, Systematic Discovery and Characterization of Regulatory Motifs in ENCODE TF Binding Experiments, *Nucleic Acids Research*, Vol. 42, No. 5, pp. 2976-2987, March, 2014.

[8] D. Quang, Y. Chen, X. Xie, DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants, *Bioinformatics*, Vol. 31, No. 5, pp. 761-763, March, 2015.

[9] G. R. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Functional Annotation of Noncoding Sequence Variants, *Nature Methods*, Vol. 11, No. 3, pp. 294-296, March, 2014.

[10] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants, *Nature Genetics*, Vol. 46, No. 3, pp. 310-315, March, 2014.

[11] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R.-S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor, *Genome Biology*, Vol. 17, No. 1, Article No. 122, June, 2016.

[12] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, J. M. Cherry, The Encyclopedia of DNA Elements (ENCODE): Data Portal Update, *Nucleic Acids Research*, Vol. 46, No. D1, pp. D794-D801, January, 2018.

[13] B. T. Lee, G. P. Barber, A. Benet-Pages, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, P. Muthuraman, L. R. Nassar, B. Nguy, T. Pereira, G. Perez, B. J. Raney, K. R. Rosenbloom, D. Schmelter, M. L. Speir, B. D. Wick, A. S. Zweig, D. Haussler, R. M. Kuhn, M. Haeussler, W. J. Kent, The UCSC Genome Browser Database: 2022 Update, *Nucleic Acids Research*, Vol. 50, No. D1, pp. D1115-D1122, January, 2022.

[14] Y. Yao, S. A. Ramsey, CERENKOV3: Clustering and Molecular Network-derived Features Improve Computational Prediction of Functional Noncoding SNPs, *Pacific Symposium on Biocomputing*, Vol. 25, pp. 535-546, 2020.

[15] S. A. Ramsey, Z. Liu, Y. Yao, B. Weeder, Combining eQTL and SNP Annotation Data to Identify Functional Noncoding SNPs in GWAS Trait-Associated Regions, in: X. Shi, X. (Eds.), *eQTL Analysis: Methods in Molecular Biology*, Vol. 2082, Humana, New York, NY, 2020, pp. 73-86.

[16] A. Buniello, J. A.-L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of Published Genome-wide Association Studies, Targeted Arrays and Summary Statistics 2019, *Nucleic Acids Research*, Vol. 47, No. D1, pp. D1005-D1012, January, 2019.

[17] The International HapMap Consortium, The International HapMap Project, *Nature*, Vol. 426, No. 6968, pp. 789-796, December, 2003.

[18] T. Beck, T. Shorter, A. J. Brookes, GWAS Central: A Comprehensive Resource for the Discovery and Comparison of Genotype and Phenotype Data from Genome-wide Association Studies, *Nucleic Acids Research*, Vol. 48, No. D1, pp. D933-D940, January, 2020.

[19] R. Huttener, L. Thorrez, T. in'tVeld, M. Granvik,

L. Snoeck, L. VanLommel, F. Schuit, GC Content of Vertebrate Exome Landscapes Reveal Areas of Accelerated Protein Evolution, *Bmc Evolutionary Biology*, Vol. 19, Article No. 144, July, 2019.

[20] R. Li, J. Q. Han, J. Liu, J. G. Zheng, R. L. Liu, A Computational Method for Prediction of rSNPs in Human Genome, *Computational Biology and Chemistry*, Vol. 62, pp. 96-103, June, 2016.

[21] R. R. Laxton, The Measure of Diversity, *Journal of Theoretical Biology*, Vol. 70, No. 1, pp. 51-67, January, 1978.

[22] N. A. Leypold, M. R. Speicher, Review: Evolutionary Conservation in Noncoding Genomic Regions, *Trends in Genetics*, Vol. 37, No. 10, pp. 903-918, October, 2021.

[23] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes, *Genome Research*, Vol. 15, No. 8, pp. 1034-1050, August, 2005.

[24] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, A. Sidow, Distribution and Intensity of Constraint in Mammalian Genomic Sequence, *Genome Research*, Vol. 15, No. 7, pp. 901-913, July, 2005.

[25] A. Pohl, M. Beato, bwtool: A Tool for BigWig Files, *Bioinformatics*, Vol. 30, No. 11, pp. 1618-1619, June, 2014.

[26] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, S. Batzoglou, Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++, *PLoS Computational Biology*, Vol. 6, No. 12, Article No. e1001025, December, 2010.

[27] A. J. Bannister, T. Kouzarides, Regulation of Chromatin by Histone Modifications, *Cell Research*, Vol. 21, No. 3, pp. 381-395, March, 2011.

[28] N. D. Heintzman, R. K. Stuart, G. Hon, Y. T. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. V. Calcar, C. X. Qu, K. A. Ching, W. Wang, Z. P. Weng, R. D. Green, G. E. Crawford, B. Ren, Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome, *Nature Genetics*, Vol. 39, No. 3, pp. 311-318, March, 2007.

[29] D. S. Gross, W. T. Garrard, Nuclease Hypersensitive Sites in Chromatin, *Annual Review of Biochemistry*, Vol. 57, pp. 159-197, January, 1988.

[30] L. Y. Song, Z. C. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B. K. Lee, N. C. Sheffield, S. Graf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Y. Wang, D. Winter, Z. Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, T. S. Furey, Open Chromatin Defined by DNaseI and FAIRE Identifies Regulatory Elements that Shape Cell-type Identity, *Genome Research*, Vol. 21, No. 10, pp. 1757-1767, October, 2011.

[31] R. Li, D. X. Zhong, R. L. Liu, H. Q. Lv, X. M. Zhang, J. Liu, J. Q. Han, A Novel Method for in Silico Identification of Regulatory SNPs in Human Genome, *Journal of Theoretical Biology*, Vol. 415, pp. 84-89, February, 2017.

[32] T. Y. Zhou, L. Yang, Y. Lu, I. Dror, A. C.-D. Machado, T. Ghane, R. DiFelice, R. Rohs, DNAshape: A Method for the High-throughput Prediction of DNA Structural Features on A Genomic Scale, *Nucleic Acids Research*, Vol. 41, No. W1, pp. W56-62, July, 2013.

[33] X. Y. Yan, J. Zhou, H. P. Huang, C. H. Wu, L. J. Sun, A. G. Song, A Weighted Range-free Localization Algorithm for Irregular Multihop Networks, *International Journal of Communication Systems*, Vol. 35, No. 10, Article No. e5153.1-e5153.16, July, 2022.

[34] X. Y. Yan, L. J. Sun, Z. X. Sun, J. Zhou, A. G. Song, Improved Hop-based Localisation Algorithm for Irregular Networks, *IET Communications*, Vol. 13, No. 5, pp. 520-527, March, 2019.

[35] J. R. Goni, A. Perez, D. Torrents, M. Orozco, Determining Promoter Location based on DNA Structure First-principles Calculations, *Genome Biology*, Vol. 8, No. 12, Article No. R263, December, 2007.

[36] J. SantaLucia, D. Hicks, The Thermodynamics of DNA Structural Motifs, *Annual Review Biophysics and Biomolecular Structure*, Vol. 33, pp. 415-440, June, 2004.

[37] L. Breiman, Bagging Predictors, *Machine Learning*, Vol. 24, No. 2, pp. 123-140, August, 1996.

[38] Z. B. Sun, Q. B. Song, X. Y. Zhu, H. L. Sun, B. W. Xu, Y. M. Zhou, A Novel Ensemble Method for Classifying Imbalanced Data, *Pattern Recognition*, Vol. 48, No. 5, pp. 1623-1637, May, 2015.

[39] G. Shafer, A Mathematical Theory of Evidence turns 40, *Internal Journal of Approximate Reasoning*, Vol. 79, pp.7-25, December, 2016.

[40] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. D. Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El-Houdaigui, T. El-Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugan, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Pilizota, I. Prosovetskaia, M. P. Sakthivel, A. I.-A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Perez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M. M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gomez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, G. R. IIsley, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, P. Flicek, Ensembl 2022, *Nucleic Acids Research*, Vol. 50, No. D1, pp. D988-D995, January, 2022.

Biographies

Rong Li received the Ph.D. degree in Control Science and Engineering from Xi'an Jiaotong University in 2017. She is currently a college lecturer at Nanjing University of Posts and Telecommunications. Her research interests mainly include artificial intelligence, data mining, and bioinformatics.

Zhi-e Lou received the Ph.D. degree in Control Theory and Control Engineering from Northeastern University (Shenyang, China) in 2018. She is currently an associate professor at Nanjing University of Posts and Telecommunications. Her research interests mainly include nonlinear system control and switching system control.