Prediction of RBPs from sequence by means of genetic algorithm and nu-SVR

**Wenjun Li Peng Yang\* and Yanyan Ge**

School of Computer Science and Engineering, Southeast University, Nanjing, China

Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China

Email: pengyang@seu.edu.cn,

**Abstract.** Post-transcriptional regulation occurs at every moment in human’s body, so it makes the identification of RNA-binding proteins (RBPs) very important, because the RBPs are indispensable accessories to post-transcriptional regulation. Although many computational methods have been developed to replace the high-costly experimental methods, most of them run slowly and the result not well enough. Based on above factors, in this study, we propose a new method namely GASVR-RBP. Firstly, we extract features from protein sequences based on physicochemical properties and Pre-in-One web server, after the feature vector space constructed, we trained eight classifiers on 9857 protein sequences with the combination of genetic algorithm (GA) and nu-SVR, and by employing the ensemble strategy, we obtained an improved performance in three test set, the accuracy are 89.3%, 84.3% and 88.8%, which higher than Naive Bayes (NB) and Random Forest (RF). These results show that our method is effective for RBPs prediction.

**Keywords:** RBPS Prediction, Feature Extract, GA, Nu-SVR.

1. Introduction

Since 1958, Crick, F.H.C. put forward the genetic central dogma, which exposes the flow of genetic information: from DNA to protein, pass through RNA. Scientists have been working on studying the relationship between the above three biological molecules for many years. RNA binding proteins (RBPs), which can interact with different RNAs to form the complex [1], RBPs are always considered as the key molecule due to its essential role in cell processes, especially in post-transcriptional regulation [2]. Hence, identifying RBPs from unknown protein sequence is a fundamental problem in biology. However, the cost of biological experiments is very high, and there have many limits. Along with the development of machine learning, more and more studies prefer to analyze the protein sequence by computational methods.

Although various computational methods have been raised to predict RBPs, most of them extract features from position specific scoring matrix (PSSM) based on evolutionary information [3], for example, Kumar et al. [4] worked out a method to predict RBPs with SVM based on evolutionary information. As we knowledge, the generation of PSSM may take a long time for RBPs prediction. Hence, it’s more convenient to extract features only from primary sequence. In order to identify the RBPs faster and accuracy, we worked out a method which only based on the sequence information [5].

Our method implements the feature selection and data classification at the same time. Firstly, we extract a 148-dimensional feature vector from protein sequence based on physicochemical properties with composition, transition and distribution (C-T-D) and conjoint triad method (CT) [6, 7], then a web server called Pre-in-One helps us to gain a 400-dimension feature vector [8]. After the feature vector space is constructed, we combine GA and nu-SVR algorithm to use GA to adjust parameters of nu-SVR. In order to avoid the deviation in single classifier, ensemble strategy is adopted to average the eight classifiers [9]. Finally, we test our model on three independent test set and realize a wonderful result.

1. Method
	1. Dataset

We use the datasets provided by Zhang, X. and Liu, S. [3]. The training set contains 2780 RBPs and 7077 non-RBPs, the RBPs are obtained from UniProt database, meanwhile the non-RBPs from PDB database. All the length of protein sequences are between 50 and 10000. After the classifiers constructed, the model will be tested on three species: i) Human dataset which contains 967 RBPs and 597 non-RBPs; ii) Saccharomyces cerevisiae (S. cerevisiae) which contains 354 RBPs and 135 non-RBPs; iii) Arabidopsis thaliana (A. thaliana) which contains 456 RBPs and 37 non-RBPs. The processing of test set is the same as that of training set.

* 1. Physicochemical Properties

We have extracted 548-dimensional vector from each protein based on amino acids sequence information. The global protein sequence descriptors C-T-D is used to extract 84 dimension features from hydrophobicity, normalized van der Waals volume, polarity and polarizability, each property can encode a 21-dimensional feature vector. A 64-dimensional vector indicates charge and polarity of side chain by CT and the rest 400 dimension is extracted by Pse-in-One server.

* + 1. *Physicochemical Properties*

We extract sequence feature based on physicochemical properties with the C-T-D and CT method, the detail process has been shown in figure 1.



 **protein sequence based on physicochemical propertiesprotein sequence based on physicochemical properties**

.

**Figure 1.** Feature extraction from protein sequence based on physicochemical properties

The C-T-D is widely used to describe global composition of amino acid sequence; Han et al. [10] developed a method using these three types of descriptors to describe the distribution of each amino acid in the protein. In table 1, we translate a protein sequence into a new sequence based on physicochemical properties, for instance, according to the hydrophobicity property, all the twenty amino acid are divided into: i) hydrophobic amino acid [C, L, V, I, M, F, W]; ii) hydrophilic amino acid [R, K, E, D, Q, N]; iii) neutral amino acids [G, A, S, T, P, H, Y]. In this case, the amino acid sequence MVVKFMDVYQR will be translated to 33313313211, based on this new sequence, we can work out that:

1. Composition descriptor:

$C\_{r}=\frac{n\_{r}}{N} \left(r=1,2,3\right) $ (1)

1. Transition descriptor:

 $T\_{rs}=T\_{rs}=\frac{n\_{rs}+n\_{sr}}{N−1}$ (2)

1. Distribution descriptor:

$D\_{rp}=\frac{POS(n\_{rp})}{N}$ (p = first, 25%, 50%, 75%, 100%) (3)

In equation (1), the variable Cr refers to the frequency of the $r$ class amino acid, in this example, proteins have been divided into three categories due to hydrophobicity property, so $r $ is from 1 to 3. The $n\_{r}$ means the number of the $r$ class amino acid in the sequence, the $N$ means the length of this sequence, equation (2) and equation (3) is the same as (1). After that, we obtain a 3-dimension feature vector based on formula (1); In formula (2), the variable $T\_{rs}$ refers to the frequency of translating from $r$ to $s$ and the opposite situation. The variable $n\_{rs}$ and $n\_{sr}$ stands for the times of the above cases respectively, for transition descriptor, we get a 3 dimension vector; In formula (3), $POS(n\_{rp})$ refers to the position of the first, 25%, 50%, 75%, 100% to the r class amino acid, there is 5 position information for every class, so we get $3∗5=15$ dimension vector for distribution descriptor. When the C-T-D program finishes, a 21-dimension feature vector is born.

**Table 1.** Distribution of amino acid according to physicochemical properties

|  |  |  |  |
| --- | --- | --- | --- |
| Physicochemical properties | The 1st class | The 2nd class | The 3rd class |
| polarity | LIFWCMVY | PATGS | HQRKNED |
| polarizability | GASDT | CPNVEQIL | MHKFRYW |
| hydrophobicity | RKEDQN | GASTPHY | CVLIMFW |
| Normalized Van der Waals volume | GASCTPD | NVEQIL | MHKFRYW |

We also employ CT to encode the charge and polarity of side chain. Firstly, all twenty amino acids are divided into 4 groups: acidic [D,E], basic [H,R,K], polar [C,G,N,Q,S,T,Y] and non-polar [A,F,I,L,M,P,V,W] [11]. We define the three successive amino acids as a unit and each unit belongs to one of the 4\*4\*4=64 dimensional vector, every value in the 64-dimensional vector stands for the normalized probability of corresponding unit [3].

*Pse-in-One*

Pse-in-one is a flexible web server with a large number of build-in properties, it can generate many feature vectors though 28 different modes for DNA, RNA and protein, it can effectively reflect the sequence’s pattern information, the server can be available at <http://bioinformatics.hitsz.ehu.cn/Pse-in-One/> Pse-in-One provides eight vector models to generate protein sequence properties, they are: 1. basic kmer; 2. auto covariance; 3. cross covariance; 4. auto-cross covariance; 5. parallel correlation pseudo amino acid composition; 6. series correlation pseudo amino acid composition; 7. general parallel correlation pseudo amino acid composition; 8. general series correlation pseudo amino acid composition. By employing the Pse-in-One server, we get a 400-dimensional feature vector for each protein sequence.

## Algorithm

### *GASV*R

GASVR is an efficient and accurate algorithm which realizes the combination of GA and nu-SVR classifier, and the process of our GASVR framework has been showed in figure 2. The nu-SVR is a model of LibSVM [12], which can be available at , it has been widely used in data classification. By setting the appropriate parameters, the SVR algorithm can achieve a reasonable result [13]. GA is a computational model for simulating the evolutionary process of natural selection and genetic mechanism based on the theory of biological evolution [14]. It starts from a random population that each single chromosome is a potential optimal solution. Through the “survival of the fittest”, chromosome will be saved. In our experiments, we use the follow fitness function to calculate the fitness:

Fitness= $a∗acc+b∗\sqrt{sen∗spec}−MSE−c∗feas−d∗SVs$ (4)

where $acc$ is accuracy of nu-SVR, $sens$ is sensitivity, $spec$ is specificity, $MSE$ is the mean square of all instances error, if the value of $MSE$ is very high, the fitness will be very low, it means the individual’s output is far away from the true class. $feas$ is the number of selected features in each individual. $SVs$ is the number of support vectors, which has been obtained in the trained nu-SVR model. After many attempts, the best parameter a=0.5, b=0.5, c=0.01, d=0.01 is used to create model.



**Figure 2.** The framework of GASVR-RBPs.

After the assessment, the optimal individual in current generation will stands out and goes straight into the next generation. The other chromosome individuals in the next generation need to be chosen from current generation according to the fitness value. Then, every generation runs the probability operation of mutation and crossover. Finally, after some iteration, we can find the global optimal solution.

In our GA framework, every chromosome is a binary sequence with a 602-bit length. The front 548 bits encoding the features of 548 dimensions extracted from the protein sequences. In this encoding, 1 in the corresponding position indicates selecting this feature, 0 represents discarding this feature. The rest 54 bits are used to dynamic tune the nu-SVR parameters: i) 14 bits for Radial Basis kernel (RBF) bandwidth gamma, the front 10 bits encodes the decimal part, and the latter 4 bits codes the integer part; ii) 20 bits for classifier parameters C, the front 10 bits encodes the decimal part, and the latter 10 bits codes the integer part; iii) 10 bits for nu-SVR parameters nu; iv) 10 bits for classification threshold which distinguishes the RBPs and non-RBPS, the range of value is between 0 and 1;

### *Ensemble GASVR-RBPs*

In order to avoid the potential error in individual classifier and maximize the performance, we adopt the ensemble strategy to combine the sequence output. All the eight classifiers are trained on the same 9857 training data set. Through averaging the results, we obtain the average output and average threshold, after the judgement, we get a better result and have a more generalization ability figure 3.



**Figure 3.** Ensemble strategy based on eight classifiers

## Performance Evaluation

Six common criteria are used to measure the performance of EnsembleGASVR-RBPs, Sensitivity (SN), Specificity (SP), Precision (PRE), Accuracy (ACC), Matthews’s correlation coefficient (MCC), they are defined as bellow:

$$SN=\frac{TP}{TP+FN} (5)$$

$$SP=\frac{TN}{TN+FP} (6)$$

$$PRE=\frac{TP}{TP+FP} (7)$$

$$ACC=\frac{TP+TN}{TP+TN+FN+FP} (8)$$

$$F−measure=\frac{2∗PRE∗SN}{PRE+SN} (9)$$

$$MCC=\frac{TP∗TN−FP∗FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} (10)$$

The RBPs are the positive set; the non-RBPS are the negative set. TP, FP, TN, FN are refer to true positive, false positive, true negative and false negative, respectively. All the indicators above are combined to show the performance of our method.

1. Result
	1. Performance of Every Generation

We spend about 24 hours to train 8 classifiers with a 4 cores 8 threads CPU, for each classifier, we repeat 20 generations and 30 chromosome individuals in every generation, after the training, we choose one of 8 classifiers, and record the performance of best member and mean performance of members of every generation, the information has been showed in figure 4, From the figure, we can see that, by the increase of iteration times, the performance of the best member and the mean performance of every generation are both increasing. Overall speaking, each generation performs better than the previous generation.



**Figure 4**. Performance of best member and mean performance of members in every generation

**Table 2.** Performance of our method with NB and RF on test data set

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Methods | SN | SP | PRE | ACC | F-measure | MCC |
| Human | EnsembleGASVR-RBPs | **0.878** | 0.916 | **0.944** | **0.893** | **0.910** | **0.781** |
| NB | 0.838 | 0.861 | 0.907 | 0.847 | 0.871 | 0.686 |
| RF | 0.667 | **0.977** | 0.907 | 0.785 | 0.793 | 0.633 |
| cerevisiae | EnsembleGASVR-RBPs | **0.848** | 0.800 | 0.917 | **0.843** | **0.881** | **0.615** |
| NB | 0.879 | 0.667 | 0.874 | 0.820 | 0.876 | 0.548 |
| RF | 0.593 | **0.919** | **0.950** | 0.683 | 0.730 | 0.460 |
| thaliana | EnsembleGASVR-RBPs | **0.884** | 0.946 | 0.995 | **0.888** | **0.936** | **0.571** |
| NB | 0.816 | 0.757 | 0.976 | 0.811 | 0.889 | 0.360 |
| RF | 0.539 | **0.973** | **0.996** | 0.572 | 0.700 | 0.270 |

* 1. Compared with Other Methods

We also compare our method with the other algorithm (NB: Naive Bayes, RF: Random Forest) on the independent dataset, the results are showed in table 2. From table 2, in Human set, EnsembleGASVR-RBPs achieves a wonderful F-measure 91.0%, which is higher than NB (87.1%), RF (79.3%), and it's the same situation in other two dataset, for ACC, our method realizes the 84.3%, higher than NB (82.0%), RF (68.3%) in S. cerevisiae dataset. On the $MCC$ index, our method is also a big lead. The reason for our advantage on index SN, SN and $PRE$ is perhaps that the other two algorithms have great error bias for RBPs prediction; they lack the ability to separate the two proteins.

Discussion

In this study, we constructed feature vector space by encoding the physicochemical properties with the C-T-D and conjoint triad method, we also employ the Pse-in-One web server. A 548-dimensional feature vector space is built on 9857 protein sequences (2780 RBPS, 7077 non-RBPs), then by combining the GA and nu-SVR classifier, we get eight classifiers, through ensemble strategy. The performance of our classifier on three test sets (human, cerevisiae, thaliana) achieves a wonderful performance compared with other algorithm.

References

1. Liu H, Wong L. Data mining tools for biological sequences[J]. *Journal of bioinformatics and computational biology*, 2003, 1(01): 139-167.
2. Joshi A, Beck Y, Michoel T. Post‐transcriptional regulatory networks play a key role in noise reduction that is conserved from micro‐organisms to mammals[J]. *The FEBS journal*, 2012, 279(18): 3501-3512.
3. Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM[J]. *Bioinformatics*, 2016, 33(6): 854-862.
4. Kumar M, Gromiha M M, Raghava G P. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information.[J]. *Journal of Molecular Recognition*, 2011, 24(2):303-313
5. Muppirala U K, Honavar V G, Dobbs D. Predicting RNA-Protein Interactions Using Only Sequence Information[J]. *BMC Bioinformatics*, 2011, 12(1):489.
6. Dehzangi A, Sohrabi S, Heffernan R, et al. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features[J]. *BMC bioinformatics*, 2015, 16(4): S1.
7. Dehzangi A, Paliwal K, Sharma A, et al. A Combination of Feature Extraction Methods with an Ensemble of Different Classifiers for Protein Structural Class Prediction Problem[J]. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 2013, 10(3):564-575.
8. Liu B, Liu F, Wang X, et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences[J]. *Nucleic Acids Research*, 2015, 43(Web Server issue):W65-W71.
9. Rapakoulia T, Theofilatos K, Kleftogiannis D, et al. EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms[J]. *Bioinformatics*, 2014, 30(16):2324-33.
10. Han L Y, Cai C Z, Lo S L, et al. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach[J]. Rna-a Publication of the Rna Society, 2004, 10(3): 355-368.
11. Cheng C W, Su E C, Hwang J K, et al. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information[J]. BMC Bioinformatics, 2008, 9(12): 1-19.
12. Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):1-27.
13. Alashwal H, Deris S, Othman R M, et al. Support vector machines for predicting protein-protein interactions using domains and hydrophobicity features[C]. International conference on computing & informatics, 2006: 1-5.
14. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2004, 33(9):1455-1465.