J Comput Aided Mol Des (2014) 28:49–60
DOI 10.1007/s10822-014-9706-6

Incorporating significant amino acid pairs and protein domains to predict RNA splicing-related proteins with functional roles

Justin Bo-Kai Hsu · Kai-Yao Huang · Tzu-Ya Weng · Chien-Hsun Huang · Tzong-Yi Lee

Received: 12 August 2013 / Accepted: 7 January 2014 / Published online: 19 January 2014

Abstract Machinery of pre-mRNA splicing is carried out through the interaction of RNA sequence elements and a variety of RNA splicing-related proteins (SRPs) (e.g. spliceosome and splicing factors). Alternative splicing, which is an important post-transcriptional regulation in eukaryotes, gives rise to multiple mature mRNA isoforms, which encodes proteins with functional diversities. However, the regulation of RNA splicing is not yet fully elucidated, partly because SRPs have not yet been exhaustively identified and the experimental identification is labor-intensive. Therefore, we are motivated to design a new method for identifying SRPs with their functional roles in the regulation of RNA splicing. The experimentally verified SRPs were manually curated from research articles. According to the functional annotation of Splicing Related Gene Database, the collected SRPs were further categorized into four functional groups including small nuclear Ribonucleoprotein, Splicing Factor, Splicing Regulation Factor and Novel Spliceosome Protein. The composition of amino acid pairs indicates that there are remarkable differences among four functional groups of SRPs. Then, support vector machines (SVMs) were utilized to learn the predictive models for identifying SRPs as well as their functional roles. The cross-validation evaluation presents that the SVM models trained with significant amino acid pairs and functional domains could provide a better predictive performance. In addition, the independent testing demonstrates that the proposed method could accurately identify SRPs in mammals/plants as well as effectively distinguish between SRPs and RNA-binding proteins. This investigation provides a practical means to identifying potential SRPs and a perspective for exploring the regulation of RNA splicing.

Keywords RNA splicing · Spliceosome · Splicing-related protein · Amino acid pair composition · Support vector machine

Justin Bo-Kai Hsu and Kai-Yao Huang have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9706-6) contains supplementary material, which is available to authorized users.

J. B.-K. Hsu
Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-chu 300, Taiwan
e-mail: justin.bokai@gmail.com

K.-Y. Huang · T.-Y. Weng · T.-Y. Lee (✉)
Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan
e-mail: francis@saturn.yzu.edu.tw

K.-Y. Huang
e-mail: kaiyao.tw@gmail.com

T.-Y. Weng
e-mail: julweng@saturn.yzu.edu.tw

C.-H. Huang
Tao-Yuan Hospital, Ministry of Health and Welfare, Taoyuan 320, Taiwan
e-mail: lithsunh@gmail.com

Introduction

The pre-mRNA splicing is required for typical eukaryotes that produce mature mRNA before it codes a correct protein through translation. The mechanism of RNA splicing is done by a series of reactions that are regulated by the splicing-related proteins (SRPs), which is a collection of small nuclear RNAs (snRNAs) and proteins recruited to

pre-mRNAs for carrying out intron excision [1, 2]. In eukaryotes, transcriptome and proteome diversities are enriched by a important mechanism of post-transcriptional regulation called alternative splicing (AS) [3]. It means a single pre-mRNA can give rise to multiple mature mRNA isoforms, which encodes proteins with different structure and function [4, 5]. Many studies have demonstrated the critical roles of AS process in various developmental stages, physiologies, diseases, and so on, and also caused higher proportion of eukaryotic genes, especially in human (higher than 50 %) [3, 6–9]. For instance, thousands of alternatively spliced transcripts in the nervous system are translated into their protein counterparts with the special capacities in learning and memory, neuronal cell recognition, neuro-transmission, ion channel function, and receptor specificity [10]. Therefore, in order to code a correct protein with flexible function, precisely excising introns from the pre-mRNA are required [11, 12].

The fidelity of intron excision is achieved by a two-step splicing. During the first step, the adenosine nucleotide located within the intron sequence known as "branch site" attacks on the 5′ splice site. The reaction generates two splicing intermediates such as a free exon 1 and a lariat-exon 2. During the second step, exon 1 attacks on the 3′ splice site and ligates to exon 2. This yields two splicing products such as a spliced exon and lariat intron [4, 12], and leads exon constitutively or alternatively spliced. Both reactions are catalyzed by the macromolecular complex, spliceosome, composed of five essential small nuclear ribonucleoprotein particles (snRNPs) and hundreds of non-snRNPs and specific proteins [1, 2]. Two types of spliceosomes have been identified in eukaryotes and are referred to as major (or U2-type) and minor (or U12-type) spliceosomes [11, 13]. For the U2-type spliceosome, each snRNP comprises of five short RNA molecular (U1, U2, U4, U5 and U6 snRNAs) bound stably to two classes of proteins, i.e., Sm proteins or Sm-like proteins, and specific proteins that are uniquely associated with only one snRNP [4, 11, 14, 15]. The function of snRNPs is useful to the spliceosome formation, which starts with U1 snRNP binding to the 5′-splice site, followed by U2 snRNP binding to the intron branch site, association of the U4/U6.U5 tri-snRNP, and release of U1 and U4 snRNPs to form the catalytic complex [1]. For the U12-type spliceosome, the U1, U2, U4, and U6 snRNAs of the major spliceosome are replaced by U11, U12, U4atac, and U6atac snRNAs, respectively. The U5 snRNA seems to function in both the major and minor spliceosome [15].

In addition to snRNPs, AS requires many of other positive or negative non-snRNP proteins as *trans*-acting factors, also known as SRPs, which are recruited to the enhancer or silencer of *cis*-acting sequence elements of the pre-mRNA to enhance or repress the regulation of splicing process by recognizing nearby splice sites [9, 16–22]. Due to the dynamic conformation and various functions of SRPs, a previous work [23] has tried to categorize SRPs into several major groups including snRNPs, splicing factors, splicing regulators and spliceosome-associated proteins. One of the well-known splicing factors is serine/arginine-rich protein (SR protein) which contains serine- and arginine-rich carboxy-terminal domains [24–26]. SR proteins are a highly conserved family of structurally and functionally related splicing factors with a dual role in splicing, affecting exons constitutively or alternatively spliced [26]. One of the major proteins in splicing regulators is hnRNP which can bind to pre-mRNA and block the binding site for splicing factors [23]. SR protein kinase is another major type of splicing regulators to modulate constitutive and AS by phosphorylating SR proteins [27]. The SRPs other than snRNPs, splicing factors and splicing regulators are classified into the category of novel spliceosome-associated proteins, due to the regulatory roles of some SRPs in splicing process could not be defined clearly so far.

With the importance of SRPs in the regulation of pre-mRNA splicing, many studies have paid attention to the proteomic analysis using mass spectrometry [9, 28–32]. In the analysis of in vitro-derived spliceosomes, 17 previously known SRPs (including hnRNP proteins) and 23 Novel SRPs are identified [33]; however, it was limited to species visible in stained 2D-gels due to the multiplicity of protein–protein and protein-RNA interactions that modulate the associations between splicing factors and pre-mRNAs. Despite more than 200 human SRPs have been identified based on comprehensive proteomic analysis over the last few years [28, 29], many of newly identified proteins have not yet been experimentally verified its function in pre-mRNA splicing [11]. Thus, Jurica and Moore [1] have manually conducted about 180 human SRPs would be premature to label those proteins as *bona fide* SRPs.

Although an increasing number of SRPs has been experimentally confirmed by mass spectrometry-based proteomic studies, the wet-lab identification is proven to be time-consuming and labor-intensive. Over the last few years, several studies have been proposed to computationally predict RNA-binding proteins (RBPs) [34, 35]. Additionally, many computational methods have been developed to identify RNA-binding residues on protein sequences [36–45]. Recently, a study has investigated the amino acid composition (AAC) in human splicing factors [46]. These published works have demonstrated their accuracy and stability; however, there is no computational method dedicated to identify SRPs with their functional roles. Thus, we are motivated to develop a systematic approach focusing on the investigation and identification of eukaryotic SRPs using the experimentally verified

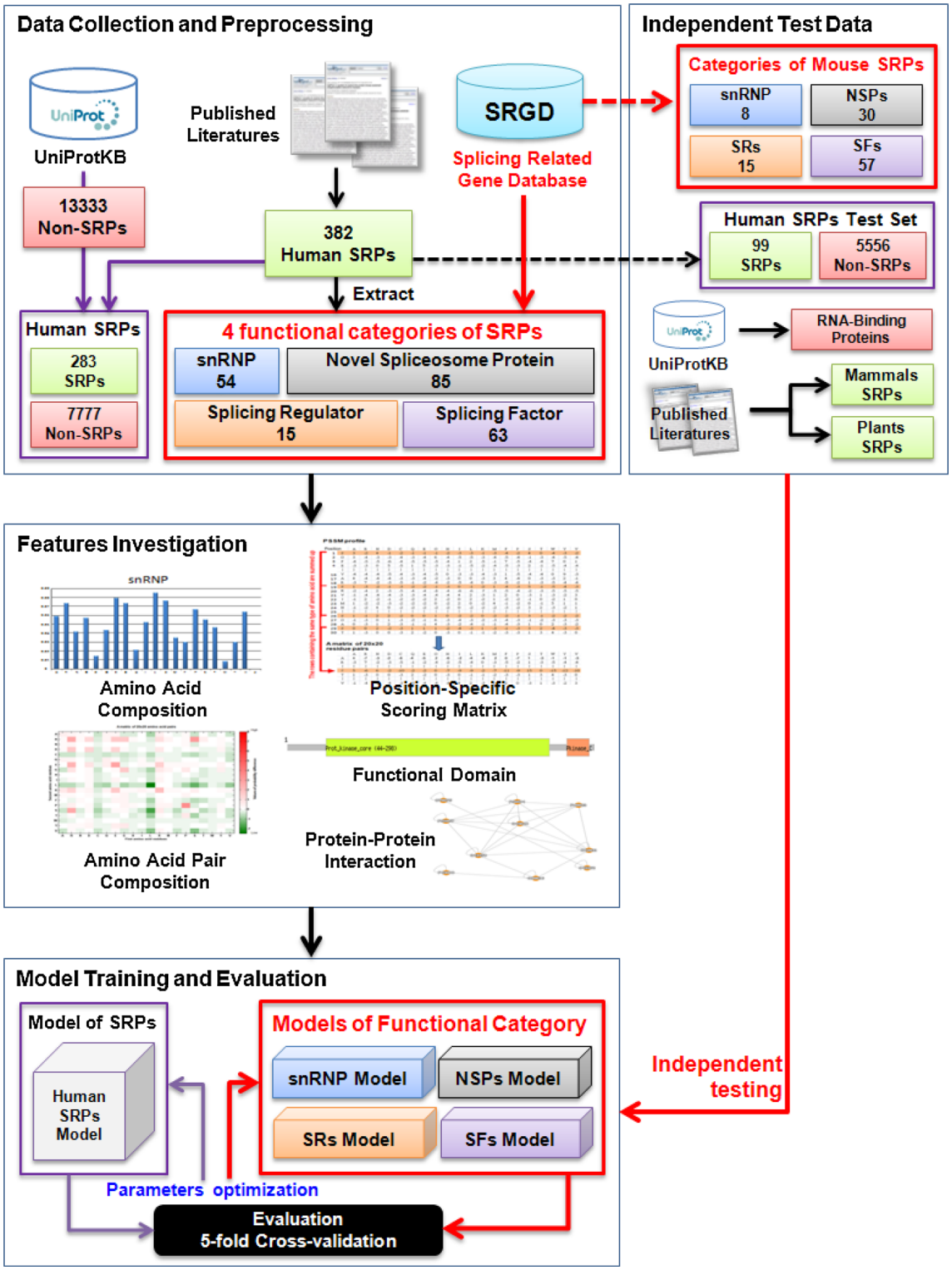

Fig. 1 The overall flowchart of the proposed method. It consists of four major parts: data collection and preprocessing, features investigation, model learning and evaluation, and independent testing

spliceosomal proteins and splicing factors. According to the functional annotations in Splicing Related Gene Database (SRGD) [47], this study further investigates into the functional roles of SRPs in RNA splicing mechanism, such as snRNP, Splicing Factor (SF), Splicing Regulation Factor (SRF) and Novel Spliceosome Protein (NSP). Furthermore,

Table 1 Data statistics after using CD-HIT

Sequence identity (%)	Positive data of training set	Positive data of independent test set	Negative data
100 (original)	283	99	19,557
90	274	94	18,897
80	271	94	18,447
70	266	94	17,727
60	249	88	16,710
50	229	82	15,255
40	217	80	13,333

the independent testing sets, which are not included in the training set, are adopted to evaluate the effectiveness of the proposed method.

Materials and methods

Data collection and preprocessing

The overall flowchart of the proposed method is depicted in Fig. 1. More than 380 experimentally confirmed SRPs in humans were manually curated from two published literatures [1, 11]. After the removal of redundant SRPs by matching to UniProtKB [48] protein entries, it resulted in 283 non-redundant SRPs which are regarded as positive data for feature investigation and model training. Additionally, human proteins which are not included in the positive data were extracted from the UniProtKB and were regarded as the candidate set of non-SRPs. In order to filter out potential noise data for non-splicing proteins, the remaining proteins consisting of keywords "RNA splicing", "spliceosome", or "splicing factors" are removed. As a result, totally 19,557 non-SRPs are regarded as negative data.

In the classification between SRPs and non-splicing proteins, the prediction performance might be overestimated due to a high sequence homology in the training set or independent test set. Thus, it is necessary to remove the homologous sequences in the dataset mentioned above, respectively. With reference to the work by Panwar et al. [49], homologous sequences in the training set are removed by CD-HIT. Firstly, CD-HIT forms a cluster with a representative sequence having the longest length which is then compared to the remaining sequences. If the similarity between a target sequence and the representative sequence is above the user-selected sequence identity threshold which refers to the pairwise sequence identity between two proteins, then the target sequence is considered homologous to the representative sequence [50]. Different values were tested for the sequence identity parameter as shown in Table 1. The resulting dataset given a sequence identity parameter of 40 % contains 217 positive sequences and 13,333 negative sequences of training set. According to the functional annotation of SRPs in SRGD, all 217 SRPs are further categorized into four major groups including snRNP, SF, SRF and NSP. This results in 54 snRNPs, 63 SFs, 15 SRFs and 85 NSPs.

Features extraction

This study aims to investigate the AAC, functional domains, and protein–protein interactions in the 217 human SRPs, as well as in the four functional groups. In order to investigating the difference of the composition of amino acids between SRPs (positive data) and non-splicing proteins (negative data), each protein sequence is represented using a vector $\{x_i,\ i = 1,\ldots,n\}$ labeled according to its corresponding protein group (e.g. splicing factor or non-splicing protein). The vector x_i has 20 elements for the AAC and 400 elements for the amino acid pair composition (AAPC). For AAC, the 20 elements specify the numbers of occurrences of 20 amino acids normalized with the total number of residues in a protein. On the other hand, for AAPC, the 400 elements specify the numbers of occurrences of 400 amino acid pairs normalized with the total number of dipeptides in a protein.

In order to identify the significant difference of amino acid pairs between positive data and negative data, a measurement of F-score [51, 52] has been applied to calculate a statistical value for each amino acid pair. The F-score of the ith value of 400 amino acid pairs is defined as:

$$\text{F}-\text{score (i)} = \frac{\left(\overline{x}_i^{(+)} - \overline{x}_i\right)^2 + \left(\overline{x}_i^{(-)} - \overline{x}_i\right)^2}{\frac{1}{n^+-1}\sum_{k=1}^{n^+}\left(x_{k,i}^{(+)} - \overline{x}_i^{(+)}\right)^2 + \frac{1}{n^--1}\sum_{k=1}^{n^-}\left(x_{k,i}^{(-)} - \overline{x}_i^{(-)}\right)^2} \quad (1)$$

where $\overline{x}_i$, $\overline{x}_i^{(+)}$ and $\overline{x}_i^{(-)}$ denote the average frequency of the ith amino acid pair in whole, positive, and negative data sets, respectively; n^+ denotes the number of positive data set and n^- denotes the number of negative data set; $x_{k,i}^{(+)}$ denotes the frequency of ith amino acid pair in the kth positive instance, and $x_{k,i}^{(-)}$ denotes the frequency of ith amino acid pair in the kth negative instance.

Several amino acid residues of a protein can go through mutation without changing its structure, and two proteins may share similar structures with different AACs. In this work, evolutionary information is obtained using position-specific scoring matrix (PSSM). PSSM profiles have been extensively utilized in protein secondary structure prediction, subcellular localization and other approaches in

bioinformatics [53–55]. The PSSM profiles of each protein were obtained by using PSI-BLAST search against the non-redundant database of protein sequences compiled by NCBI [56]. Due to the fact that the data consists of protein sequences with variable length, a weighted score of features is obtained by summing up the position-specific scores of the same amino acids occurring in a protein sequence to get a uniform number of features. Figure S1 (Additional File 1) displays in detail how to generate a 400-dimensional (20 × 20 residue pairs) PSSM feature vector for each SRP and non-splicing protein. PSSM profile is a matrix of $m \times 20$ elements where m represents the protein sequence length and 20 represents the position-specific scores for each type of amino acid. Then, the PSSM profile is transformed to a 20 × 20 matrix by summing up each row of same amino acid in the PSSM profile and the variable is denoted as "x". Finally, every element of 400-dimensional PSSM vector is divided by the length of the sequence and then is scaled by $\frac{1}{1+e^{-x}}$ for normalizing the values between 0 and 1.

Previous works on protein prediction have exhibited the ability of distinguishable domain regions in the classification of proteins [57]. In this work, domain information is regarded as a feature for classifying the functional roles of SRPs. To investigate the preference of functional domains in each functional group, this study referred to the annotations in InterPro [58]. InterPro is an integrated resource, which was developed initially as a means of rationalizing the complementary efforts of the PROSITE [59], PRINTS [60], Pfam [61], and ProDom [62] databases, for providing protein "signatures" such as protein families, domains and functional sites. The domain information of each SRP in the training data is collected by referring to its corresponding InterPro ID in the UniProtKB database. The collected domains are then utilized to evaluate the predictive performance in identifying SRPs as well as the functional roles of SRPs.

Model learning and cross-validation evaluation

Support vector machine (SVM) is applied to generate computational models that incorporate the encoded set of features. Based on binary classification, the concept of SVM is to map the input samples into a higher dimensional space using a kernel function, and then to find a hyperplane that discriminates between the two classes with maximal margin and minimal error. A public SVM library, LibSVM [63], is used to train the predictive model with positive and negative training sets, which are encoded with reference to various training features. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$, where S_i and S_j are the feature vectors of two input samples, is selected as the kernel function of SVM. The gamma parameter (γ) defines the extent of the influence of a single training sample, with low values meaning higher influence [64]. Cross-validation is important to the application of the predictor [65]. The predictive performance of the constructed models is evaluated by performing k-fold cross-validation. The training data is divided into k groups by splitting each dataset into k approximately equal sized subgroups. In this work, k is set to five. During cross-validation, each subgroup is regarded as the validation set in turn, and the remainder is regarded as the training set. Next, the following measures of predictive performance of the trained models are defined:

$$\text{Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{3}$$

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \tag{4}$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively. Additionally, the usefulness values of true positive (UsefulnessPOS) and true negative (UsefulnessNEG) predictions for independent testing are defined as well:

$$\text{UsefulnessPOS} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{UsefulnessNEG} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{6}$$

A hybrid approach is employed in this work by combining different sets of feature vectors with the goal of improving prediction performance. Three AAC-based types of hybrid combinations are explored. In the first combination, the effect of combining AAC with the information of functional domain is explored. In the second combination, the effect of combining amino acid pairs composition with the functional domain is explored. In the third combination, the effect of combining PSSM with functional domain is explored. Additionally, the parameters of the predictive model, cost and gamma value of the SVM models are optimized to maximize predictive accuracy. In optimization of SVM parameter C and RBF kernel parameter gamma, the grid search is applied to obtain the parameters that achieve the best accuracy during k-fold cross-validation. Then, the hybrid combinations of features that yield the highest accuracy are employed to construct predictive models for independent testing. Finally, the SVM model trained with the combined features and the selected parameters (C and gamma) are evaluated the predictive performance using independent testing data.

Independent testing

In case of over-fitting to the training set the performance of constructed models might be overestimated in classifying between SRPs and non-splicing proteins. Hence, the independent testing is required to evaluate the actual performance of the predictive models. The SRPs of independent testing set is constructed from UniProtKB by extracting the human proteins which are not among the positive data of training set and are obtained from the resulting dataset by collecting protein entries annotated as "RNA splicing", "spliceosome", or "splicing factors". The UniProtKB uses such annotations to define a protein entry that has been experimentally identified to be essential for RNA splicing. In order to filter out potential noise data for non-splicing proteins, the remaining proteins consisting of keyword "RNA-binding" are removed. This yielded 99 protein sequences which are then regarded as positive data for independent testing. Given a sequence identity parameter of 40 % in CD-HIT, the resulting dataset contains 80 positive sequences, as presented in Table 1. To generate the negative data for independent testing, the negative data (13,333 protein sequences) is then randomly divided into two sets: 7,777 protein sequences are regarded as negative data for model training and 5,556 protein sequences are regarded as negative data for independent testing.

In order to test the effectiveness of the proposed method in identifying SRPs from other mammalian species, a total of 142 experimentally verified splicing factors in mouse and rat species were extracted from published literature [11]. Additionally, 309 SRPs of *Arabidopsis thaliana* were collected from the ASRG database [66]. Furthermore, to test the ability of our method in differentiating between SRPs and RBPs, the human proteins consisting of keyword "RNA-binding" and not included in the positive data of training set are extracted from UniProtKB. This resulted in a total of 584 human RBPs for the evaluation of predictive specificity.

Results and discussion

Composition of amino acids in splicing-related proteins

The comparison of AAC between SRPs and non-splicing proteins is presented in Fig. 2, which indicates the enrichments of Arginine (R), Aspartic Acid (D), Glutamic Acid (E) and Lysine (K) residues in SRPs. The dominance of these amino acid residues indicates its contribution in RBP and protein–protein interactions. The over-representation of R and K in SRPs is reasonable because these positively charged residules can easily interact with negatively charged RNA. Another abundant amino acid group observed in SRPs is the negatively charged residue (D and E) which are easily located on surface area of a protein for interacting with other SRPs. Interestingly, Leucine (L) is

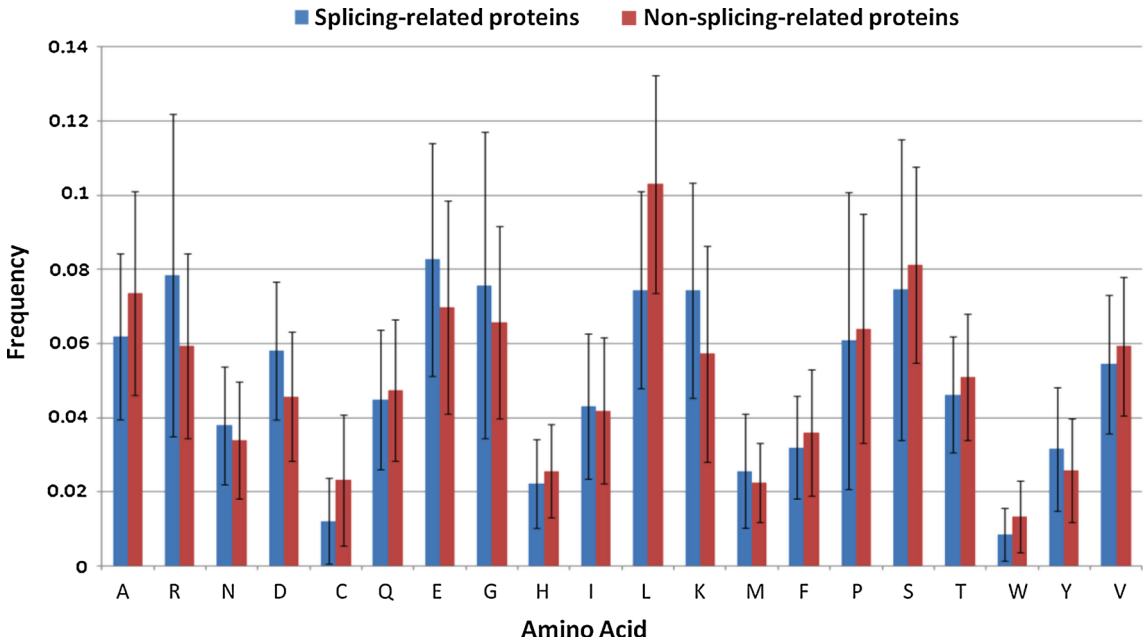

Fig. 2 Percent composition of twenty amino acids between splicing-related proteins (positive data) and non-splicing proteins (negative data). This investigation indicates the enrichments of Arginine (R), Aspartic Acid (D), Glutamic Acid (E) and Lysine (K) residues in SRPs. The abundance of R and K in splicing factors is reasonable because these positively charged residules can easily interact with negatively charged RNA. Another abundant amino acid group observed in splicing factors is D and E which are negatively charged residues and are easily located on the surface area of a protein to interact with other splicing factors

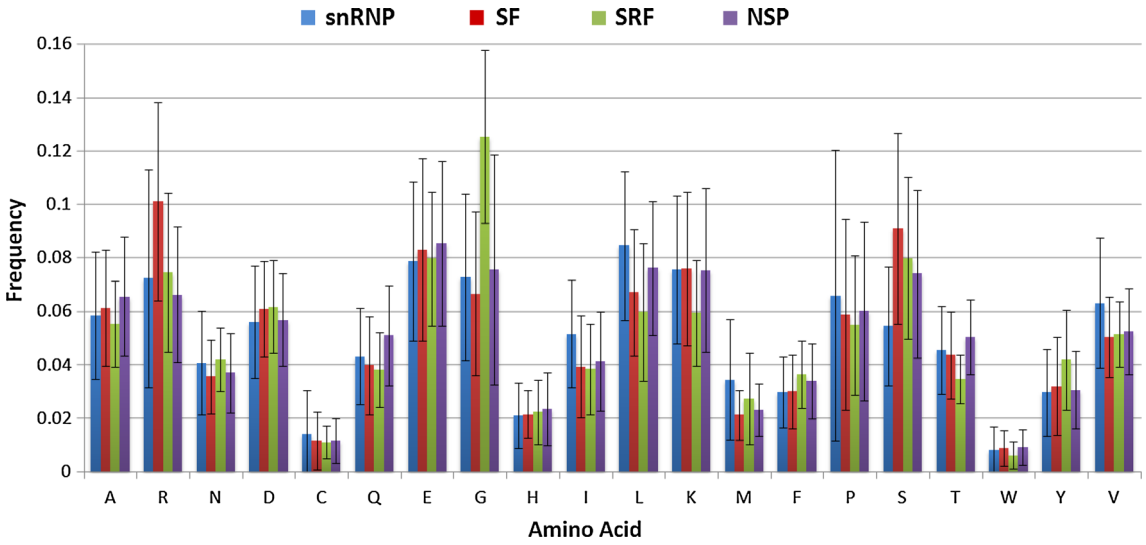

Fig. 3 Percent composition of twenty amino acids in four functional groups of SRPs

observed to be the most prominent among all under-represented residues.

To investigate the difference of AAC among four functional groups, the percent composition of 20 amino acids in each group is illustrated in Fig. 3. In snRNP group, an over-represented amino acid group is nonpolar, aliphatic residues including Isoleucine (I), L, Methionine (M) and Valine (V). In the group of splicing factors (SFs), a positively charged residue (R) is over-represented and a polar residue (S) is slightly enriched. A remarkable enrichment of Glycine (G) is observed in SRF group. The small size and flexibility of G residues is probably making it suitable for the structural adjustments required during the splicing regulation [67]. In NSP group, there is no over- or under-representation of amino acids when comparing to other three groups. Another view of AAPC may identify the difference between NSP and other groups.

Previous studies have demonstrated that AAPC could be a useful feature for yielding a better performance as compared to AAC-based methods [52, 68, 69]. In order to investigate this claim in terms of identifying SRPs, the frequency differences of 400 amino acid pairs between 217 positive and 7,777 negative sequences are calculated, as shown in Figure S2 (Additional File 1). In the 20 × 20 matrix, amino acid pairs marked in red indicates over-representation in SRPs while amino acid pairs marked in green indicates under-representation. This investigation shows that DD pair is over-represented in SRPs as well as D residues paired with K, E and R. It would be noticed that Cysteine (C) residues paired with other residues are under-represented in SRPs. In an attempt to make the comparison of AAPC among snRNP, SF, SRF and NSP groups, furthermore, the frequency differences of 400 amino acid pairs of four functional groups are presented in Fig. 4. This investigation shows that SF and SRF groups contain remarkable enrichments of amino acid pairs. Many amino acid pairs are slightly enriched in snRNP group. For NSP group, there is no significant amino acid pair when comparing to other three groups.

In the prediction of SRPs as well as the functional roles, the importance of amino acid pairs is further evaluated by means of F-score measurement for identifying the significant amino acid pairs in a specific data set. The positive and negative frequencies of each amino acid pair are computed by means of dividing the number of positive or negative proteins containing the target amino acid pair by the total number of positive or negative sequences, respectively. Then, the statistical significance of each amino acid pair is calculated by the hypergeometric test (p value) [70]:

$$P(t) = \sum_{t}^{T} \frac{C_t^T \cdot C_{k-t}^{K-T}}{C_k^K} \tag{7}$$

where K is the background set represented by the number of all proteins and T is the sample set represented by the number of SRPs; k is the number of all proteins having the target amino acid pair and t is the number of SRPs containing the target amino acid pair. A smaller p value stands for a greater statistical significance. In this work, the amino acid pair containing a p value <0.05 is considered as a statistically significant amino acid pair (SSAAP).

Investigation of functional domains in splicing-related proteins

In order to analyze preference of functional domains in SRPs, the experimentally verified domains of 217 SRPs in

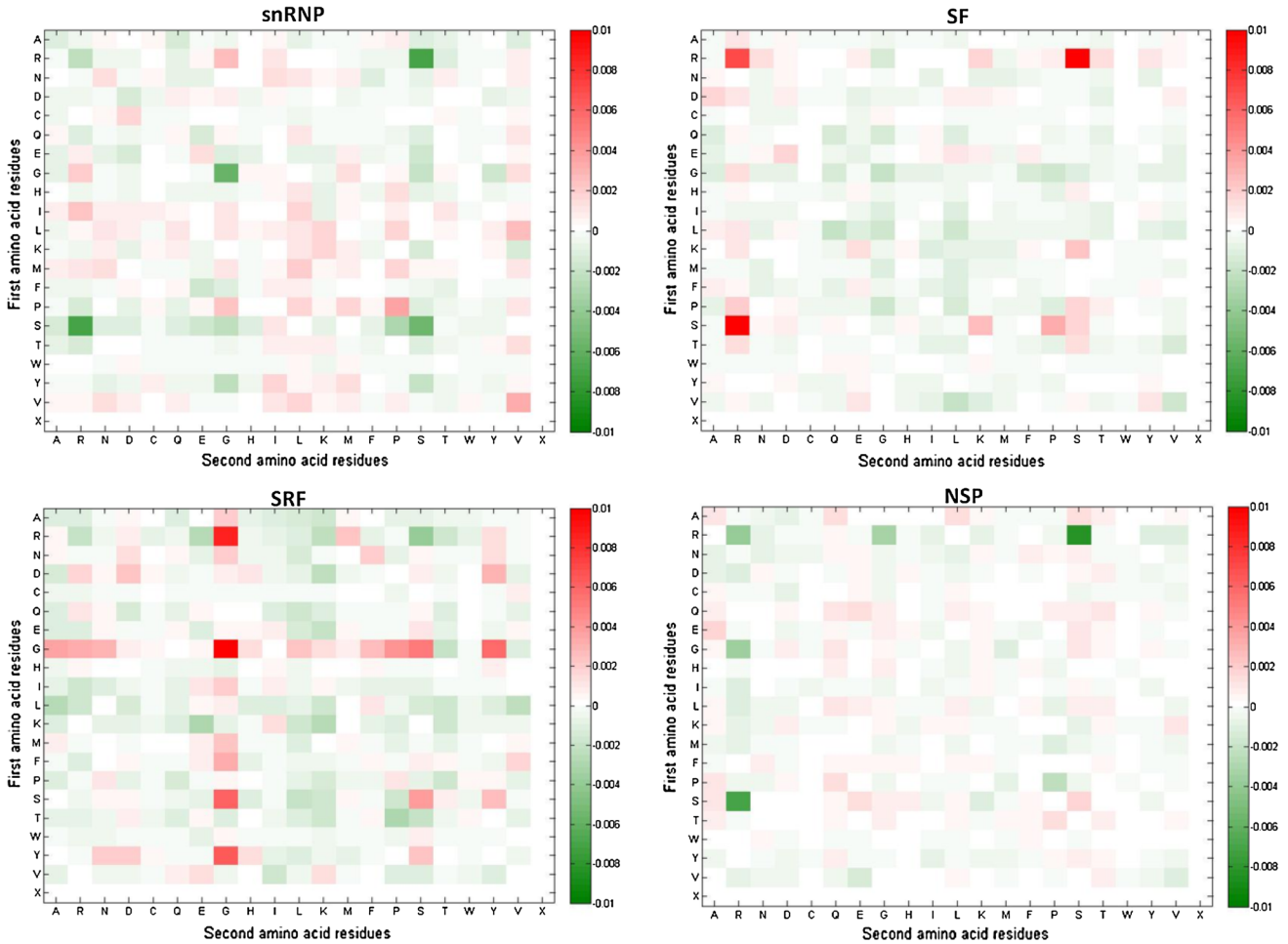

Fig. 4 The frequency differences of 20×20 amino acid pairs among snRNP, SF, SRF and NSP. The amino acid pair with *red box* indicates an over-representation in a specific functional group comparing to the other three groups; on the other hand, *green box* means an under-representation

the training data is collected by referring to the "InterPro" field in UniProtKB. This resulted to a total of 284 functional domains existing in SRPs. As shown in Table S1 (Additional File 1), the functional domains which are present in more than 5 SRPs are selected as distinguishable domains in the classification between SRPs and non-splicing proteins. It is observed that the most enriched functional domain is the "Nucleotide-bd a/b plait" with InterPro ID: IPR012677 which exists in 48 SRPs. Another enrichment of functional domain is the "RNA recognition motif domain" with InterPro ID: IPR000504 which exists in 46 SRPs. Additionally, the preference of InterPro domains in four functional groups is also investigated. As shown in Table S2 (Additional File 1), a total of 104, 99, 21 and 151 InterPro domains are detected in snRNPs, SFs, SRFs and NSPs, respectively. Table 2 shows the top ten enriched protein domains in snRNP, SF, SRF and NSP groups. The most distinguishable functional domain in snRNPs is "Like-Sm ribonucleoprotein (LSM) domain" with InterPro ID: IPR001163. The "Nucleotide-bd a/b plait" and "RNA recognition motif domain" are the most enriched domains in SF, SRF and NSP groups. Although the most enriched domains are common among SR, SRF and NSP groups, several domains in SRF group, such as "Protein kinase-like domain", "Protein kinase, ATP binding site", "Protein kinase, catalytic domain" and "Zinc finger, CHHC-type", are distinguishable from SR and NSP groups. In order to evaluate the predictive performance of using the functional domains, the SVM models are trained using a x-dimensional vector, where x is the number of the distinguishable domains,represented by a binary score: 1 if present and 0 otherwise.

Cross-validation performance in the prediction of SRPs

In the binary classification between 217 SRPs and 7,777 non-splicing proteins, the SVM models trained with four basic features such as AAC, AAPC, PSSM and functional

Table 2 Top ten enriched InterPro protein domains in four functional groups of SRPs

Ranking	InterPro ID	InterPro description	Number of SRPs
snRNP (54 SRPs)			
1	IPR001163	Like-Sm ribonucleoprotein (LSM) domain	15
2	IPR010920	Like-Sm ribonucleoprotein (LSM)-related domain	15
3	IPR000504	RNA recognition motif domain	8
4	IPR012677	Nucleotide-binding, alpha–beta plait	8
5	IPR015880	Zinc finger, C2H2-like	4
6	IPR003604	Zinc finger, U1-type	3
7	IPR000690	Zinc finger, C2H2-type matrin	3
8	IPR024888	U1 small nuclear ribonucleoprotein A/U2 small nuclear ribonucleoprotein B	2
9	IPR013085	Zinc finger, U1-C type	2
10	IPR015943	WD40/YVTN repeat-like-containing domain	2
SF (63 SRPs)			
1	IPR012677	Nucleotide-binding, alpha–beta plait	23
2	IPR000504	RNA recognition motif domain	22
3	IPR014001	DEAD-like helicase	5
4	IPR001650	Helicase, C-terminal	5
5	IPR011545	DNA/RNA helicase, DEAD/DEAH box type, N-terminal	5
6	IPR011046	WD40 repeat-like-containing domain	5
7	IPR015943	WD40/YVTN repeat-like-containing domain	5
8	IPR019775	WD40 repeat, conserved site	5
9	IPR001680	WD40 repeat	5
10	IPR017986	WD40-repeat-containing domain	5
SRF (15 SRPs)			
1	IPR012677	Nucleotide-binding, alpha–beta plait	8
2	IPR000504	RNA recognition motif domain	8
3	IPR011009	Protein kinase-like domain	3
4	IPR017441	Protein kinase, ATP binding site	3
5	IPR008271	Serine/threonine-protein kinase, active site	3
6	IPR000719	Protein kinase, catalytic domain	3
7	IPR012996	Zinc finger, CHHC-type	2
8	IPR002290	Serine/threonine-/dual-specificity protein kinase, catalytic domain	1
9	IPR011989	Armadillo-like helical	1
10	IPR016024	Armadillo-type fold	1
NSP (85 SRPs)			
1	IPR000504	RNA recognition motif domain	18
2	IPR012677	Nucleotide-binding, alpha–beta plait	18
3	IPR014001	DEAD-like helicase	7
4	IPR001650	Helicase, C-terminal	7
5	IPR011545	DNA/RNA helicase, DEAD/DEAH box type, N-terminal	7
6	IPR011046	WD40 repeat-like-containing domain	5
7	IPR015943	WD40/YVTN repeat-like-containing domain	5
8	IPR001680	WD40 repeat	5
9	IPR017986	WD40-repeat-containing domain	5
10	IPR000467	D111/G-patch	5

Table 3 Cross-validation performance of the investigated features in differentiating between 217 SRPs and 7,777 non-splicing proteins

Features	Sensitivity (%)	Specificity (%)	Accuracy (%)
Amino acid composition (AAC)	76.0	77.8	77.8
Amino acid pair composition (AAPC)	79.2	78.0	78.1
Statistically significant amino acid pairs (SSAAPs)	78.3	78.6	78.6
Position-specific scoring matrix (PSSM)	79.7	79.9	79.9
Functional domain (FD)	44.7	90.0	88.9
AAC + FD	74.7	81.0	80.8
SSAAPs + FD	79.2	81.7	81.6
PSSM + FD	80.6	82.0	82.0
SSAAPs + PSSM + FD	**80.6**	**82.6**	**82.5**

The best performance is marked in bold

domain (FD) are evaluated the predictive performance using five-fold cross-validation. Additionally, the SSAAPs detected by F-score measurement are also utilized to examine the improvement of prediction performance when comparing to the SVM model trained with AAPC. As presented in Table 3, among the basic features, the SVM model trained with functional domain (FD model) yields the best accuracy (88.9 %) but gives a worst sensitivity. The SSAAPs model could provide a better prediction accuracy than the SVM model using all of 400 amino acid pairs. With the consideration of a balanced sensitivity and specificity, the PSSM model outperforms other SVM models. In the cross-validation evaluation of three AAC-based combinations, the SVM model trained with the hybrid combination of PSSM and FD yields the best performance. In our further test, the SVM model learned from

Table 4 Cross-validation performance of the investigated features in categorizing 217 SRPs into four functional groups

Features	snRNP (54 sequences)			SF (63 sequences)			SRF (15 sequences)			NSP (85 sequences)		
	Sn (%)	Sp (%)	Acc (%)	Sn (%)	Sp (%)	Acc (%)	Sn (%)	Sp (%)	Acc (%)	Sn (%)	Sp (%)	Acc (%)
Amino acid composition (AAC)	70.4	69.9	70.0	69.8	67.5	68.2	73.3	73.3	73.3	61.2	60.6	60.8
Amino acid pair composition (AAPC)	70.4	71.20	70.9	68.3	66.2	66.8	80.0	74.3	74.7	62.3	62.1	62.2
Statistically significant amino acid pairs (SSAAPs)	74.1	71.8	72.4	71.4	71.4	71.4	80.0	76.7	77.0	64.7	66.7	65.9
Position-specific scoring matrix (PSSM)	68.5	66.9	67.3	63.4	64.3	64.1	80.0	79.2	79.3	60.0	59.8	59.9
Functional domain (FD)	77.7	78.5	78.3	76.2	77.9	77.4	**86.7**	**84.2**	**84.3**	77.6	77.3	77.4
AAC + FD	74.1	73.0	73.3	73.0	74.7	74.2	80.0	79.2	79.3	71.8	70.5	71.0
SSAAPs + FD	**81.5**	**81.6**	**81.6**	**81.0**	**81.2**	**81.1**	86.7	80.7	81.1	**78.8**	**79.5**	**79.3**
PSSM + FD	72.2	74.8	74.2	69.8	70.8	70.5	86.7	80.2	80.6	72.9	72.0	72.4
SSAAPs + PSSM + FD	79.6	79.1	79.3	79.3	80.5	80.2	**86.7**	**84.2**	**84.3**	78.8	78.8	78.8

The best performance is marked in bold

the combination of SSAAPs, PSSM and FD could provide a best and balanced performance with 80.6 % sensitivity, 82.6 % specificity, and 82.5 % accuracy.

Cross-validation performance in classifying the functional roles of SRPs

In the multi-class classification among four functional groups of 217 SRPs, the one-against-all SVM is adopted. As given in Table 4, when distinguishing 54 snRNPs from 163 SRPs, the SVM model trained with SSAAPs and FD achieves a best performance with 81.5 % sensitivity, 81.6 % specificity and 81.6 % accuracy. In the differentiation between 63 SFs and other 154 SRPs, the SVM model trained with SSAAPs and FD could provide a best performance with 81.0 % sensitivity, 81.2 % specificity and 81.1 % accuracy. In the identification of 15 SRFs, the FD model gives a best performance with 86.7 % sensitivity, 84.2 % specificity and 84.3 % accuracy, which is equal to the preidction ability of SVM model learned from the SSAAPs, PSSM and FD. In the classification of 85 NSPs, the SVM model trained with SSAAPs and FD performs best in comparison to other models. Interestingly, the models trained with PSSM could not perform as better in the prediction of SRPs, partly due to the limited number of data set in the four functional groups. Overall, the SVM models trained with the combination of SSAAPs and FD could provide an average accuracy of over 80.0 % in the classification among 54 snRNPs, 63 SFs, 15 SRFs and 85 NSPs.

Independent testing performance

After the cross-validation evaluation, the SVM models containing best performance are further examined by independent testing data. As shown in Table 5, the SVM model trained with the combination of SSAAPs, PSSM and FD could provide a sensitivity of 88.8 % in positive testing data (80 potential SRPs) and a specificity of 84.3 % in negative testing data (5,556 non-splicing proteins) in human. In order to test the ability of the selected model to identify SRPs from other mammalian species, a total of 142 experimentally verified SRPs in mouse and rat were manually extracted from published literature and were used to test the best model learned from human SRPs. Table 5 shows that the human model could yield a sensitivity of

Table 5 The independent testing performance

Dataset	Human SRPs	Mammal SRPs (mouse and rat)	Plant SRPs (*A. thaliana*)	Human RBPs
Number of positive data	80	142	309	–
Number of negative data	5,556	–	–	584
True positive (TP)	71	135	252	–
False negative (FN)	9	7	57	–
True negative (TN)	4,684	–	–	501
False positive (FP)	872	–	–	83
Sensitivity (%)	88.8	95.0	81.6	–
Specificity (%)	84.3	–	–	85.8
UsefulnessPOS (%)	7.5	–	–	–
UsefulnessNEG (%)	99.8	–	–	–

RBPs RNA-binding proteins

Table 6 Independent testing performance of cross-classification among four functional groups of 110 non-homologous SRPs in mouse

Functional group	Best features	Number of positive data	Number of negative data	Sn (%)	Sp (%)	Acc (%)	UPos (%)	UNeg (%)
snRNP	SSAAPs + FD	8	102	75	74.5	74.5	18.8	97.4
SF	SSAAPs + FD	57	53	70.2	73.6	71.8	74.1	69.6
SRF	FD	15	95	86.7	81.1	81.7	41.9	97.5
NSP	SSAAPs + FD	30	80	73.3	73.7	73.6	51.2	88.1

UPos usefulnessPOS, *UNeg* usefulnessNEG

95.0 % on mouse and rat SRPs. Additionally, the independent testing shows that the selected model could correctly identify 81.6 % of 309 *Arabidopsis thaliana* SRPs. In order to evaluate the ability of the selected model in differentiating between SRPs and RBPs, the independent testing demonstrates that the selected model could achieve a specificity of 85.8 % in 584 human RBPs extracted from UniProtKB. Moreover, a total of 110 mouse SRPs with their functional roles, which are non-homologous to human SRPs, are used to examine the selected models in classifying the functional roles of human SRPs. As given in Table 6, the selected models with best combination of hybrid features could yield the accuracies of 74.5, 71.8, 81.7 and 73.6 % in classifying 8 snRNPs, 57 SFs, 15 SRFs and 30 NSPs, respectively.

Conclusions

In this work, the investigation of AAC reveals that there is a remarkable enrichment of amino acids in SRPs, as well as in snRNP, SF and SRF groups. The investigation of AAPC also reveals that there are significant amino acid pairs in SRPs and remarkable frequency differences of amino acid pairs among four functional groups. The preference of functional domains in SRPs is also investigated and utilized to identify the SRPs with their functional roles. The evaluations of cross-validation and independent testing show that the SVM model trained with the information of SSAAPs, PSSM and FD could provide a promising performance. Although the feature of PSSM is not effecitive in classifying the functional roles of SRPs, the model trained with only FD could provide a favorable performance. Overall, the SVM models trained with the combination of SSAAPs and FD could provide an effective classification among snRNPs, SFs, SRFs and NSPs. Moreover, the independent testing indicates that the proposed method could identify the SRPs in mammals and plant. Recently, various computational methods have been proposed for the identification of RBPs. The independent testing also demonstrates that the selected model could distinguish the SRPs from RBPs.

The importance of SRPs have been exhibited in pre-mRNA splicing as well as the transcript diversity of AS in several diseases. However, in vivo or in vitro identification of SRPs are subject to technical limitations. Thus, the selected models with best performance could be adopted to implement a web-based tool for identifying the SRPs with their functional roles. In eukaryotic cells SRPs work together to carry out their functional roles in RNA splicing mechanism by protein–protein interactions (PPIs) and protein-RNA interactions. With an attempt to study the protein interaction network among SRPs, the information of experimentally verified protein–protein interactions is integrated from five public databases, as shown in Table S3 (Additional File 1). Table S4 (Additional File 1) shows that a high percentage of SRPs are involved in protein–protein interactions among themselves. This preliminary analysis indicates that a protein participating in many PPIs with SRPs might be a potential SRP or plays an important role in the regulation of RNA splicing. Thus, the consideration of PPIs between SRPs and other proteins could be a practical means to identifying novel SRPs in prospective study.

Acknowledgments The authors would like to sincerely thank the National Science Council of the Republic of China for financially supporting this research under Contract No. 101-2628-E-155-002-MY2 and 102-2221-E-155-069.

References

1. Jurica MS, Moore MJ (2003) Mol Cell 12:5
2. Zahler AM, Lane WS, Stolk JA, Roth MB (1992) Genes Dev 6:837
3. Keren H, Lev-Maor G, Ast G (2010) Nat Rev Genet 11:345
4. Hui JY (2009) Sci China Ser C Life Sci 52:253
5. Hsu JBK, Bretana NA, Lee TY, Huang HD (2011) Plos One 6:e27567
6. Wang ET, Sandberg R, Luo SJ, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Nature 456:470
7. Johnson JM, Castle J, Garrett-Engele P, Kan ZY, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Science 302:2141
8. Chen L, Zheng SK (2009) Genome Biol 10:R3
9. Ben-Dov C, Hartmann B, Lundgren J, Valcarcel J (2008) J Biol Chem 283:1229
10. Grabowski PJ, Black DL (2001) Prog Neurobiol 65:289

11. Barbosa-Morais NL, Carmo-Fonseca M, Aparicio S (2006) Genome Res 16:66
12. Reed R (2000) Curr Opin Cell Biol 12:340
13. Patel AA, Steitz JA (2003) Nat Rev Mol Cell Biol 4:960
14. Johnson PJ (2002) Proc Natl Acad Sci USA 99:3359
15. Wahl MC, Will CL, Luhrmann R (2009) Cell 136:701
16. Stamm S, Ben-Ari S, Rafalska I, Tang YS, Zhang ZY, Toiber D, Thanaraj TA, Soreq H (2005) Gene 344:1
17. Matlin AJ, Clark F, Smith CWJ (2005) Nat Rev Mol Cell Biol 6:386
18. Cartegni L, Chew SL, Krainer AR (2002) Nat Rev Genet 3:285
19. Maniatis T, Tasic B (2002) Nature 418:236
20. Smith CWJ, Valcarcel J (2000) Trends Biochem Sci 25:381
21. Black DL (2003) Ann Rev Biochem 72:291
22. Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y (2010) Nucleic Acids Res 38:W281
23. Wang BB, Brendel V (2004) Genome Biol 5:R102
24. Mueller WF, Hertel KJ (2011) Landes Bioscience and Springer Science+Business Media
25. Long JC, Caceres JF (2009) Biochem J 417:15
26. Cazalla D, Newton K, Caceres JF (2005) Mol Cell Biol 25:2969
27. Stojdl DF, Bell JC (1999) Biochem Cell Biol 77:293
28. Zhou ZL, Licklider LJ, Gygi SP, Reed R (2002) Nature 419:182
29. Rappsilber J, Ryder U, Lamond AI, Mann M (2002) Genome Res 12:1231
30. Kasyapa CS, Kunapuli P, Cowell JK (2005) Exp Cell Res 309:78
31. Chen YIG, Moore RE, Ge HY, Young MK, Lee TD, Stevens SW (2007) Nucleic Acids Res 35:3928
32. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome Res 18:1381
33. Neubauer G, King A, Rappsilber J, Calvio C, Watson M, Ajuh P, Sleeman J, Lamond A, Mann M (1998) Nat Genet 20:46
34. Kumar M, Gromiha MM, Raghava GP (2010) J Mol Recognit 24:303
35. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ (2004) RNA 10:355
36. Ma X, Guo J, Wu J, Liu H, Yu J, Xie J, Sun X (2011) Proteins 79:1230
37. Wang L, Huang C, Yang MQ, Yang JY (2010) BMC Syst Biol 4(Suppl 1):S3
38. Murakami Y, Spriggs RV, Nakamura H, Jones S (2010) Nucleic Acids Res 38:W412
39. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L (2010) Bioinformatics 26:1616
40. Maetschke SR, Yuan Z (2009) BMC Bioinforma 10:341
41. Wang Y, Xue Z, Shen G, Xu J (2008) Amino Acids 35:295
42. Tong J, Jiang P, Lu ZH (2008) Comput Methods Programs Biomed 90:148
43. Kumar M, Gromiha MM, Raghava GP (2008) Proteins 71:189
44. Wang L, Brown SJ (2006) Conf Proc IEEE Eng Med Biol Soc 1:5830
45. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D (2006) RNA 12:1450
46. Hsu JB, Bretana NA, Lee TY, Huang HD (2011) PLoS One 6:e27567
47. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) Nucleic Acids Res 36:D959
48. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) Nucleic Acids Res 32:D115
49. Panwar B, Raghava GP (2010) BMC Genom 11:507
50. Li W, Jaroszewski L, Godzik A (2001) Bioinformatics 17:282
51. Lin C.-J, Chen Y.-W (2003) NIPS 2003 feature selection challenge 1
52. Chen SA, Lee TY, Ou YY (2010) BMC Bioinforma 11:536
53. Jones DT (1999) J Mol Biol 292:195
54. Xie D, Li A, Wang MH, Fan ZW, Feng HQ (2005) Nucleic Acids Res 33:W105
55. Ou YY, Gromiha MM, Chen SA, Suwa M (2008) Comput Biol Chem 32:227
56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389
57. Wang L, Huang C, Yang JY (2011) BMC Genom 11(Suppl 3):S2
58. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) Nucleic Acids Res 37:D211
59. Bairoch A (1991) Nucleic Acids Res 19(Suppl):2241
60. Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ (1994) Nucleic Acids Res 22:3590
61. Sonnhammer EL, Eddy SR, Durbin R (1997) Proteins 28:405
62. Corpet F, Gouzy J, Kahn D (1998) Nucleic Acids Res 26:323
63. Chang C.-C, Lin C.-J (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)
64. Peng H, Ozaki T, Haggan-Ozaki V, Toyoda Y (2003) IEEE Trans Neural Netw 14:432
65. Chou KC, Shen HB (2007) Anal Biochem 370:1
66. Wang BB, Brendel V (2004) Genome Biol 5:R102
67. Kumar M, Gromiha AM, Raghava GPS (2008) Proteins Struct Funct Bioinforma 71:189
68. Bhasin M, Raghava GP (2004) J Biol Chem 279:23262
69. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK (2007) Nucleic Acids Res 35:W588
70. Sadygov RG, Yates JR 3rd (2003) Anal Chem 75:3792