Genomics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides

Bharat Panwar [1], Gajendra P.S. Raghava *

Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A, Chandigarh, India

ARTICLE INFO

Article history:
Received 2 April 2014
Accepted 23 January 2015
Available online xxxx

Keywords:
Protein-interacting nucleotide (PIN)
Binary profile of patterns (BPP)
Tri-nucleotide composition profile of patterns (TNCPP)
SVM
Prediction
RNApin

ABSTRACT

The RNA–protein interactions play a diverse role in the cells, thus identification of RNA–protein interface is essential for the biologist to understand their function. In the past, several methods have been developed for predicting RNA interacting residues in proteins, but limited efforts have been made for the identification of protein-interacting nucleotides in RNAs. In order to discriminate protein-interacting and non-interacting nucleotides, we used various classifiers (NaiveBayes, NaiveBayesMultinomial, BayesNet, ComplementNaiveBayes, MultilayerPerceptron, J48, SMO, RandomForest, SMO and SVM^{light}) for prediction model development using various features and achieved highest 83.92% sensitivity, 84.82 specificity, 84.62% accuracy and 0.62 Matthew's correlation coefficient by SVM^{light} based models. We observed that certain tri-nucleotides like ACA, ACC, AGA, CAC, CCA, GAG, UGA, and UUU preferred in protein-interaction. All the models have been developed using a non-redundant dataset and are evaluated using five-fold cross validation technique. A web-server called RNApin has been developed for the scientific community (http://crdd.osdd.net/raghava/rnapin/).

1. Introduction

The interaction of RNA molecules and RNA-binding proteins (RBPs) play diverse roles in cells including protein translation, gene expression and regulation [1]. There is a large amount of RNA present in every cell, but these RBPs are selectively bound to the particular RNA at a specific site [2,3,4]. Role of RNA–protein interactions is well established for the complete functionality of cell, and in the case of its failure, this leads to the various human genetic diseases [5] such as fragile X syndrome [6], paraneoplastic neurologic syndromes [7], spinal muscular atrophy [8], myotonic dystrophy [9] and fragile X tremor ataxia syndrome [10].

Detection of protein interacting nucleotides is important to understand the underlying mechanism of RNA–protein interaction. X-ray crystal structure determination of RNA–protein complexes is a common practice to detect PINs in RNA but structural availability of these complexes is very low in comparison to total protein interacting RNAs. There are several other experimental techniques such as RNA EMSA [11], SELEX (systemic evolution of ligands exponential enrichment), CLIP [12], pull-down assay, oligonucleotide-targeted RNase H protection assays [13], RIP-ChIP [14], ribonomics [15] and Ribotrap [16] available for the detection of protein binding RNAs. These techniques are expensive, laborious and unable to provide exact information of protein interacting nucleotides. In the past, several methods have been developed for the prediction of RNA interacting residues (amino acids) in the protein sequences [17,18,19,20] but limited efforts have been made for predicting protein interacting nucleotides (PINs) in the RNA sequences [21]. Therefore, there is an urgent need to develop computational tool for this problem.

Recently, many studies have suggested important steps to develop any biological prediction method [22,23,24,25,26,27,28]. In this study a systematic attempt has been made to develop in silico tool for the prediction of PINs in RNA sequences. We analyzed the patterns of both protein interacting and non-interacting nucleotides and found that significant differences were present. A machine learning technique 'support vector machine' has been applied. We used different binary and compositional approaches and achieved highest 0.62 MCC and 0.889 AUC by tri-nucleotide composition profile of patterns (TNCPP) approach. In order to provide service to the global scientific community, this TNCPP based prediction model has been implemented in the form of a web-server called *RNApin*.

2. Material and methods

2.1. Datasets

We retrieved a total of 1546 protein-interacting RNA chains (RNA-1546) of PDB from PRIDB database [29]. We used these RNA chains and created 25% non-redundant 'RNA-208' dataset of 208 RNA chains using BLASTCLUST software. We considered only RNA chains having

* Corresponding author at: Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India. Fax: +91 172 2690632, 2690585.
E-mail addresses: bharat@imtech.res.in (B. Panwar), raghava@imtech.res.in (G.P.S. Raghava).
URL: http://www.imtech.res.in/raghava/ (G.P.S. Raghava).
[1] Present address: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

http://dx.doi.org/10.1016/j.ygeno.2015.01.005
0888-7543/

length of more than 10 nucleotides. Furthermore, we assigned each nucleotide of these RNA chains into protein interacting and non-interacting nucleotides using cutoff distance of 5.0 cutoff Å. It means that if the distance between nucleotide and any amino acid of protein chains was less or equal than 5.0 Å, then nucleotide assigned as protein interacting otherwise is assigned as non-interacting. In this way, we assigned a total of 46,582 nucleotides of RNA-208 dataset into the 10,198 protein interacting and 36,384 non-interacting nucleotides. We used 5.0 Å as cut-off because this contains almost all different kind of interactions and mostly used in the past for the prediction of RNA-interacting amino acids [30].

2.2. Creation of sliding window

In the past, sliding (overlapping) window based strategy has been applied in various residue/nucleotide level prediction methods [20,31,17]. In this study, we also created sliding window patterns of different 3–25 lengths from RNA-208 dataset. If the central nucleotide of the window was protein-interacting then whole window pattern was considered as positive pattern otherwise considered as negative patterns [32]. To generate fixed length window size of terminal nucleotides, we added a dummy 'X' nucleotide at both terminals of each RNA chain. The number of dummy nucleotides was calculated with $(L - 1) / 2$ formula (where L is the length of the pattern). It means that each nucleotide of RNA-208 dataset was once used at the central position of window pattern. Finally, we created a total of 10,198 positive and 36,384 negative patterns.

2.3. Binary profile of patterns

The numerical representation of window patterns is necessary for the machine learning tools, and BPP based strategy is one of the widely adopted approach for the window-based machine learning [31]. In BPP approach, we represented A, C, G, U and X nucleotides of all window patterns in the binary form of {1,0,0,0,0}, {0,1,0,0,0}, {0,0,1,0,0}, {0,0,0,1,0} and {0,0,0,0,1} respectively. BPP generated five times higher input features than the window size (e.g. 19-nucleotide long window pattern generates a total of 95 (19×5) input features). These binary representations of window pattern give information of nucleotide availability at a specific position during machine learning based prediction model development.

2.4. Composition profile of patterns

The composition of window pattern can also be used as input feature of machine learning [33,34].

2.4.1. Mono-nucleotide composition profile of patterns

In MNCPP, we calculated mono-nucleotide composition of all nucleotides (A, C, G, U and X) for each window pattern separately. These five numerical values of composition were used as SVM input.

2.4.2. Di-nucleotide composition profile of patterns

In DNCPP, the di-nucleotide (AA, AC, AG, CG, AU,…, XX) composition of each window pattern was calculated separately. It provided a total of 25 numerical values, which were used as SVM input. The DNCPP approach has advancement over MNCPP and that it also provides information of neighboring nucleotides.

2.4.3. Tri-nucleotide composition profile of patterns

In TNCPP, we calculated tri-nucleotide (AAA, AAC, AAG,…, XXX) composition of each window pattern separately. For each window pattern, we found a total of 125 numerical values, which were used as SVM input features.

2.5. Support vector machine

In this study, a machine learning technique, support vector machine (SVM) was applied, which is based on the structural risk minimization principle of statistics learning theory. SVMs are a set of related supervised learning methods used for classification and regression mode [35]. It has options of different parameters and kernels (e.g. linear, polynomial, radial basis function and sigmoidal) to optimize according to need. We implemented SVM^{light} Version 6.02 package [36] of SVM and machine learning. We applied various parameters and three different (linear, polynomial and radial basis function) kernels to develop different prediction models.

2.6. WEKA

WEKA is a single package and platform of different classifier [37]. We applied WEKA 3.6.6 version, which consists of different classifiers such as NaiveBayes, NaiveBayesMultinomial, BayesNet, ComplementNaiveBayes, MultilayerPerceptron, J48, SMO, RandomForest and SMO. We have used all these machine learning algorithms for the development of different prediction models.

2.7. Five-fold cross validation

The validation of the model is an important step for the development of any prediction method. There are several techniques available for validation of any prediction models like jack-knife test or leave-one-out cross validation (LOOCV), n-fold cross validation etc. [38]. Although, jackknife or LOOCV cross-validation is the most objective and consistent, but it is a time-consuming process especially for the residue-level prediction [17,25,39]. In this study, we used widely accepted five-fold cross-validation technique for training, testing and evaluation of SVM models [40,41]. In this process, first we divided all positive and negative window patterns into five parts randomly. Each of these five sets consists of one-fifth of total positive and one-fifth of total negative window patterns. In five-fold cross validation technique, we used four sets as training and the remaining one set as testing. This process was repeated five times in such a way that each set was used once as a test set. We calculated performance of each test set and overall performance of the prediction model is an overall performance of these five test sets.

2.8. Evaluation parameters

We used various evaluation parameters such as sensitivity (Eq. (1)), specificity (Eq. (2)), accuracy (Eq. (3)) and MCC (Eq. (4)) values for evaluating prediction models [42]:

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (3)$$

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}} \quad (4)$$

where *TP*, *TN*, *FP* and *FN* are True Positives, True Negative, False Positives and False Negatives respectively.

The above-mentioned parameters are threshold-dependent; therefore, we also calculated threshold-independent evaluation parameter, AUC (Area Under Curve) values for each prediction model in the ROC (Receiver Operating Curve) plots. The RNApin web-server provides prediction results by calculating probability score for each nucleotide of the given RNA sequence. In order to present SVM score effectively, we

Please cite this article as: B. Panwar, G.P.S. Raghava, Genomics (2015), http://dx.doi.org/10.1016/j.ygeno.2015.01.005

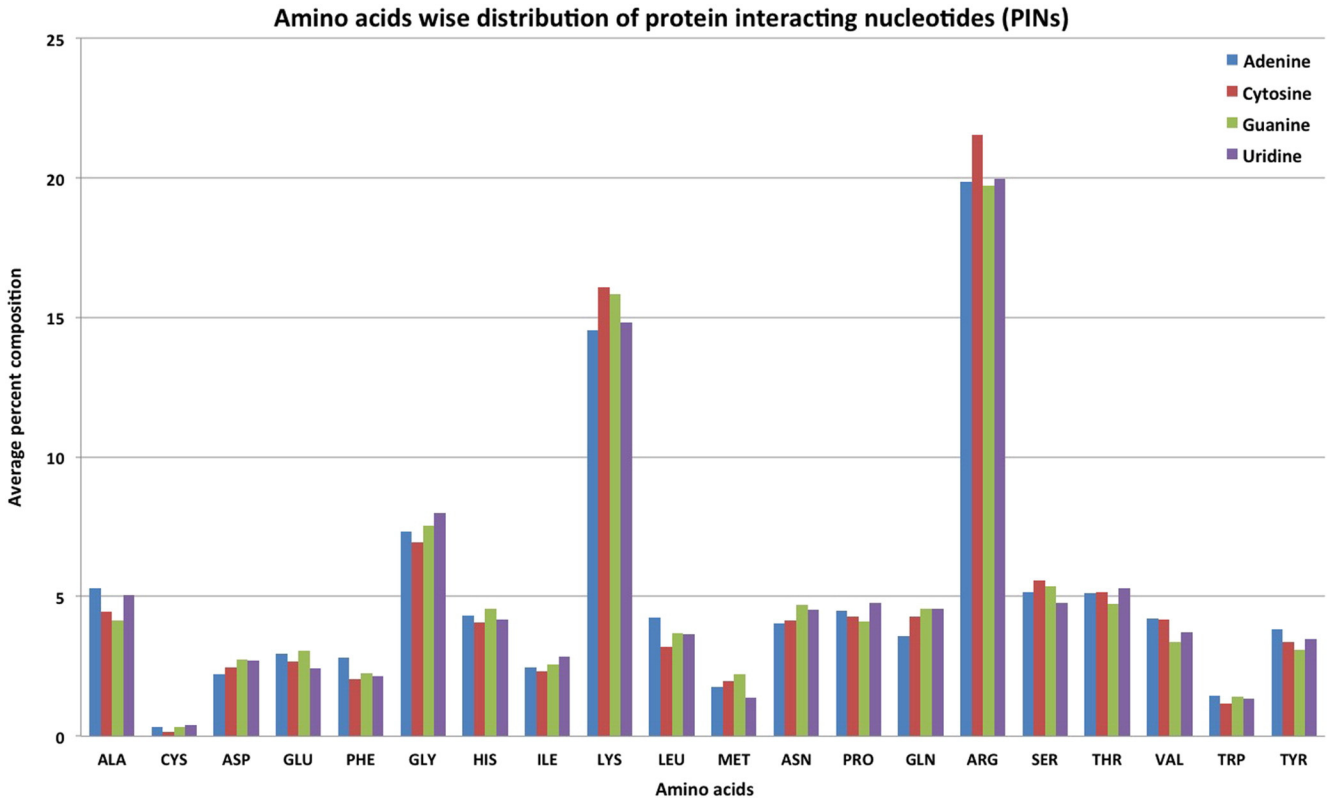

Fig. 1. Amino acids wise composition of protein interacting nucleotides.

calculated probability score using Eq. (5). We present score for display single digit between 0 to 9 and called probability score in this study. Depending on model SVM score vary between 2.5 (around) and −2.5 (around). We converted SVM score into probability score using the following steps. First, all SVM scores of more than 1.5 were assigned 1.5 and likewise less than −1.5 were assigned as −1.5. This way all score falls between −1.5 to 1.5. Secondly, 1.5 is added to each score so score falls between 0.0 to 3.0. In order to keep the number between 0 to 9, we divide each number by 3.0 and multiplied by 9. The following equation is used for computing the probability score.

$$Probability\ \ score = \mathrm{int}\left(\frac{SVM\,score\ +\ 1.5}{3} \times\ 9\right) \tag{5}$$

We used probability score in RNApin webserver instead of simple SVM score for each nucleotide because it is easy to display with every

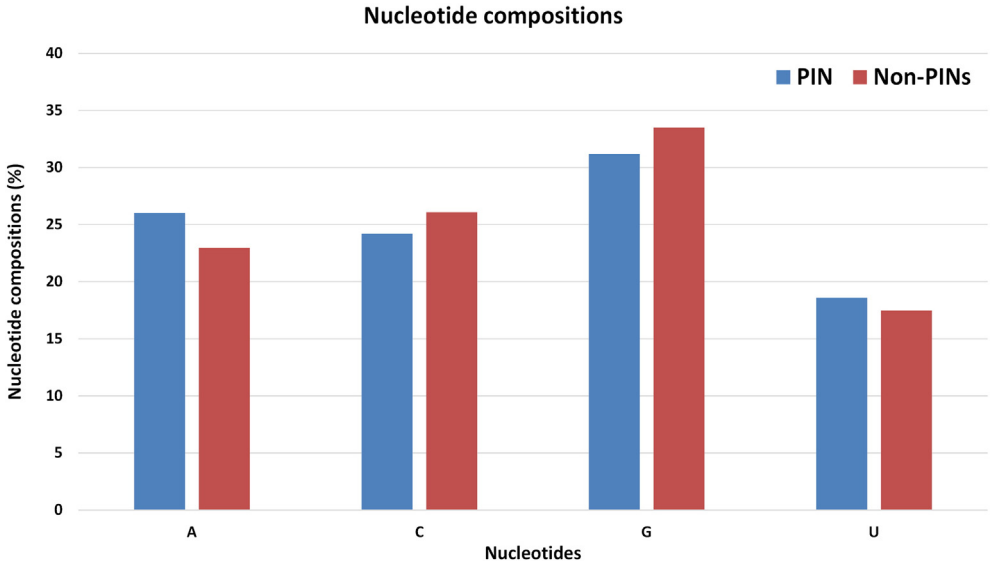

Fig. 2. Compositions of protein-interacting and non-interacting nucleotides.

Please cite this article as: B. Panwar, G.P.S. Raghava, Genomics (2015), http://dx.doi.org/10.1016/j.ygeno.2015.01.005

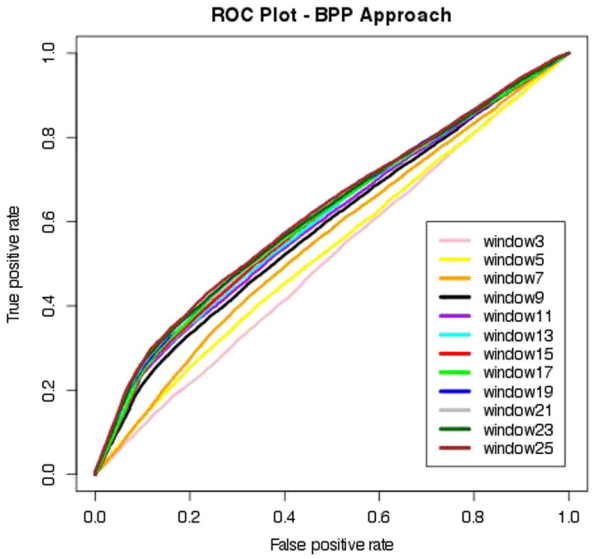

Fig. 3. ROC graph showing prediction performances of BPP approach based SVM models for 3–25 window sizes.

nucleotide of RNA sequence. The probability scores range from 0–9, where scores of 0–4 and 5–9 predicted as non-protein interacting and protein-interacting nucleotides respectively (at default 0.0 threshold level).

3. Results

3.1. Analysis of protein interacting nucleotides

Initially we extracted a total of 1546 protein-interacting RNA chains from PRIDB database [29] and used for our preliminary analysis. We analyzed amino acids wise interaction preference of different nucleotides and found that arginine, lysine and glycine were three most preferred whereas cysteine, methionine and tryptophan were non-preferred nucleotide interacting amino acids (Fig. 1). These observations agree with previous studies on RNA–protein complexes [17,43]. We also compute preference of residues with different nucleotides but no preference was observed. Moreover, we analyzed length-wise variation of a total of 35,063 protein interacting nucleotide stretches in RNA-1546 dataset. Most of the interacting stretches (92.88%) were 1–15 nucleotides long but single, di and tri-nucleotides were most abundant. Earlier Gromiha et al. showed that 17%, 15%, 15%, 16% and 11% of binding stretches are accommodate with mono, di, tri, tetra and penta nucleotides respectively [43] whereas we found that 17%, 17%, 16%, 12%, and 9% of stretches are constituted by mono, di, tri, tetra and penta-nucleotides respectively (Supplementary Fig. S1).

We calculated nucleotide compositions and found that there is no significant difference between the protein interacting and non-interacting nucleotides (Fig. 2); these observations agree with the previous study [43]. We calculated the composition of all possible pair of di-nucleotides (Supplementary Fig. S2) and tri-nucleotides (Supplementary Fig. S3) and found that there are some differences present in the compositions.

In the past, various residue/nucleotide level prediction methods have been developed on the basis of overlapping (sliding) window pattern strategy [20,31,17]. In this, we created overlapping window patterns of different 3–25 lengths from RNA-208 dataset. A pattern is assigned protein interacting or positive if the nucleotide at its center is protein interacting otherwise it was assigned as negative or non-protein interacting. To discriminate these positive and negative patterns, we applied various approaches and developed different SVM based prediction models. All the models have been evaluated using five-fold cross validation technique.

3.2. Performance of binary profile of patterns

In the past, various biological prediction methods have been developed using binary profile of patterns (BPP) approach [31,44]. Therefore, we created binary profiles of positive and negative patterns of different window sizes (see Material and methods section). These BPPs were used as input for the SVM based machine learning. Different kernels and parameters were optimized, but prediction performance was not good. We achieved maximum 61.57% sensitivity, 54.89% specificity,

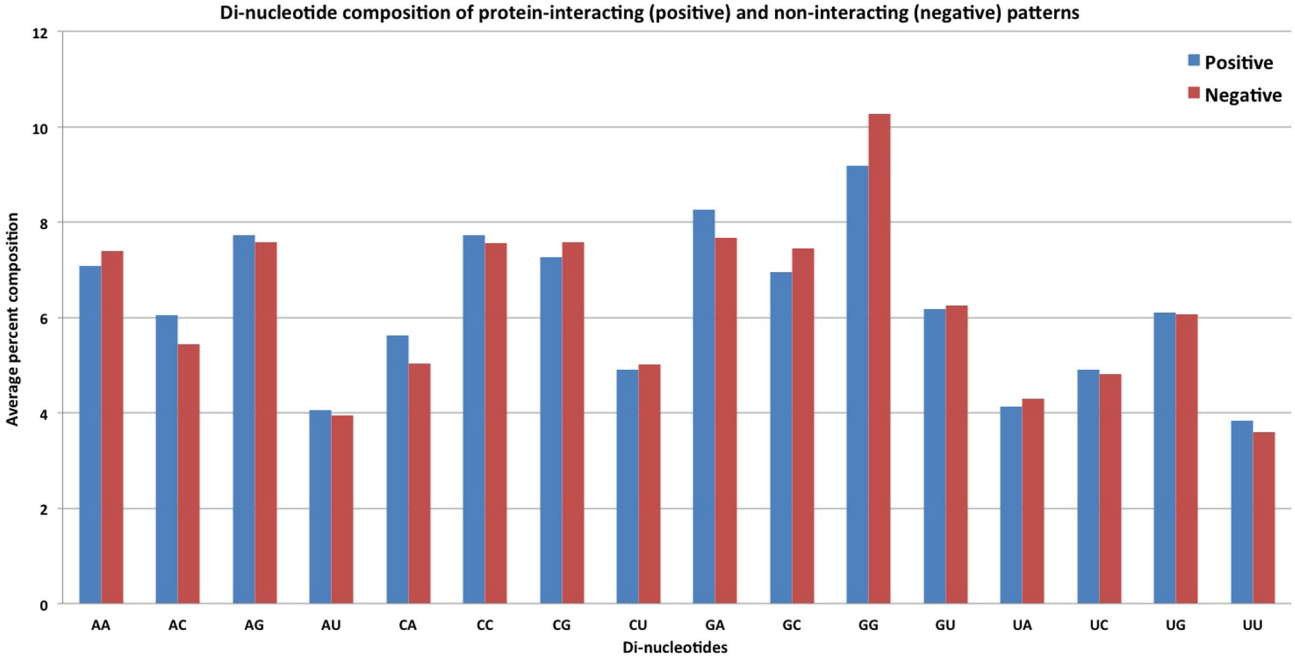

Fig. 4. Di-nucleotide composition of protein interacting (positive) and non-interacting (negative) patterns.

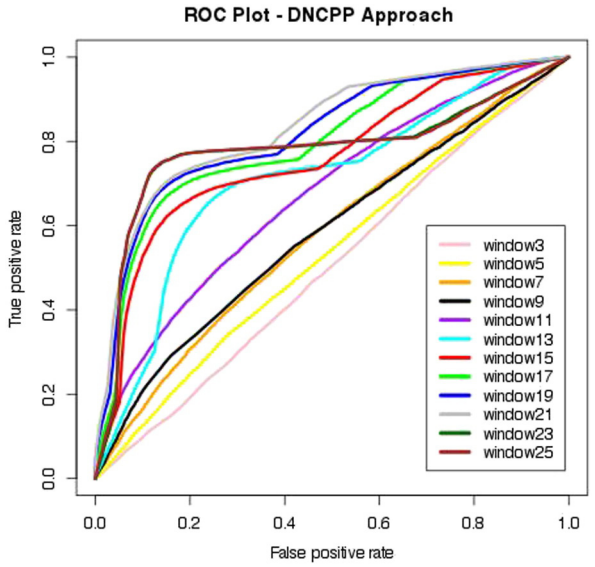

Fig. 5. ROC graph showing prediction performances of DNCPP approach based SVM models for 3–25 window sizes.

56.35% accuracy, 0.13 MCC and 0.622 AUC for window length of 25 (Fig. 3, Supplementary Table S1).

3.3. Performance of composition profile of patterns

The composition profile of patterns (CPP) can also be used for the prediction model development, when residue/nucleotide based compositional differences present between the positive and negative patterns [34,33].

3.3.1. Mono-nucleotide composition profile of patterns (MNCPP)

In the past, various prediction methods have been developed using nucleotide/amino acid composition based approach [31,45]. We calculated nucleotide compositions of the protein-interacting (positive) and non-interacting (negative) patterns and observed that there was no nucleotide-wise preference for protein-interaction (Supplementary Fig. S4). We used these positive and negative composition profiles as input for the SVM based machine learning. As expected, all the performances were very poor and achieved only 53.69% sensitivity, 52.51% specificity, 52.76% accuracy, 0.05 MCC and 0.564 AUC for window length of 21 (Supplementary Fig. S5, Supplementary Table S2).

3.3.2. Di-nucleotide composition profile of patterns (DNCPP)

Simple mono-nucleotides composition provides information of nucleotide fraction in each pattern, whereas di-nucleotide composition provides fraction information as well as the order and neighboring nucleotide information. In DNCPP approach, we calculated di-nucleotide composition of all positive and negative patterns (Fig. 4) and used these DNCPP as SVM input. We achieved maximum 74.81% sensitivity, 76.72% specificity, 76.31% accuracy, 0.45 MCC and 0.832 AUC for window length of 21 (Fig. 5, Supplementary Table S3).

3.3.3. Tri-nucleotide composition profile of patterns (TNCPP)

Tri-nucleotide composition is more informative than di-nucleotide composition because it provides information of two neighboring nucleotides. In TNCPP approach, we calculated tri-nucleotide composition of all positive and negative patterns (Fig. 6). These TNCPPs were used as an input for SVM based machine learning. We optimized different kernels and parameters on all window sizes (3–25) and finally selected the best performing prediction model. We achieved highest 83.92% sensitivity, 84.82% specificity, 84.62% accuracy, 0.62 MCC and 0.889 AUC for window length of 19 (Fig. 7, Supplementary Table S4).

We also tried different classifiers such as NaiveBayes, NaiveBayesMultinomial, BayesNet, ComplementNaiveBayes, MultilayerPerceptron, J48, SMO, RandomForest and SMO using WEKA and achieved 0.07, 0.09, 0.10, 0.10, 0.16, 0.38, 0.46, 0.47 and 0.52 values of MCC respectively (Table 1). It means SVM^{light} based models achieved highest 0.62 MCC for predicting protein-interacting nucleotides in the RNA sequences.

4. Discussion

The RNA–protein interactions are involved in various biological processes. In order to understand and investigate those interactions, it is

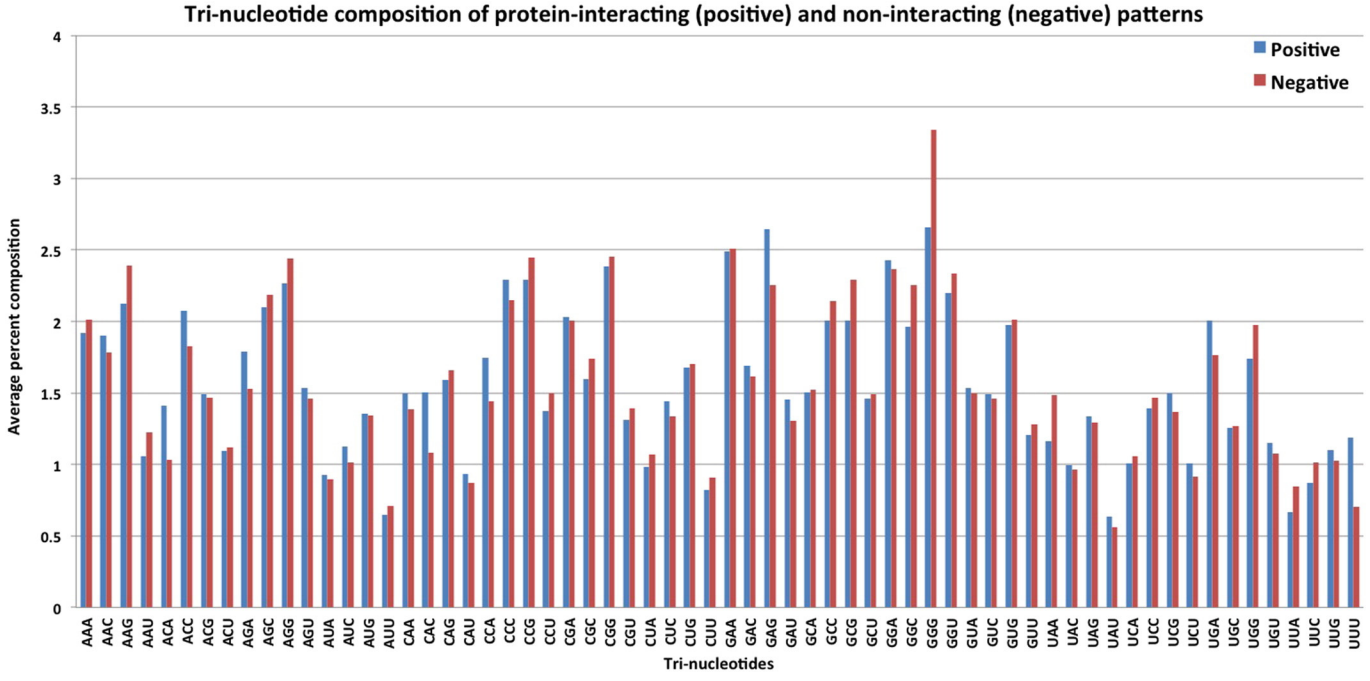

Fig. 6. Tri-nucleotide composition of protein interacting (positive) and non-interacting (negative) patterns.

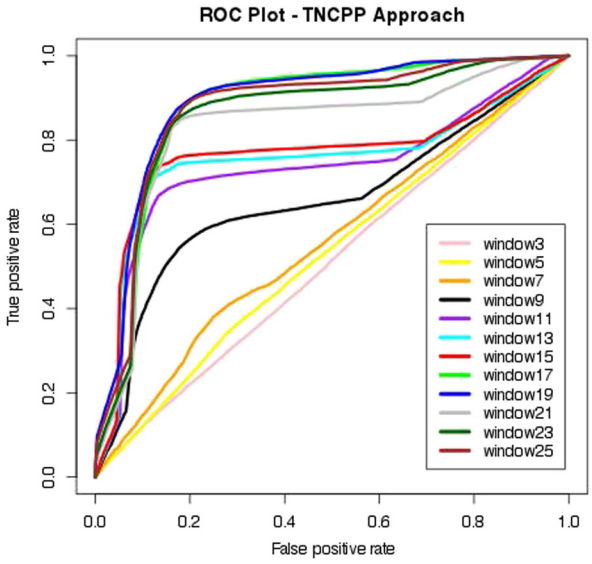

Fig. 7. ROC graph showing prediction performances of TNCPP approach based SVM models for 3–25 window sizes.

important to identify interacting amino acids and nucleotides. There are several prediction methods that have been developed to predict RNA-interacting amino acids in the protein sequences but limited method available for the prediction of interacting nucleotides in RNA sequence. We created RNA-208 dataset of 208 RNA chains from PRIDB database [29]. We calculated the RNA interaction preference of each amino acid and found that Arg, Lys and Gly are most abundant RNA-interacting nucleotides. Thereafter, interaction preference of each nucleotide for every amino acid was calculated and observed that there was no nucleotide-wise preference present or very little preference present e.g. cytosine slightly preferred to interact with arginine in comparison to the other nucleotides (Fig. 1). It was interesting to analyze the length of protein interacting nucleotide stretch and found that most of the interacting stretches are 1–15 nucleotides long, where single, di and tri-nucleotides were most abundant (Supplementary Fig. S1).

To develop a prediction tool, it is important to convert biological knowledge/information into the machine-readable numerical forms. In the past, several studies have used sliding window-based strategy to develop residue/nucleotide level prediction [20,17]. We created sliding window of different length and assigned window pattern as positive if the central nucleotide of the window was protein interacting otherwise assigned as negative. This assignment provided a total of 10,198 positive and 36,384 negative window patterns. The next challenge is how to discriminate these positive and negative patterns efficiently. The BPP is a widely used approach for this task [44,31]. This approach provides position-wise nucleotide information of the window pattern. Therefore, we applied BPP approach, but it achieved maximum 61.57% sensitivity, 54.89% specificity, 56.35% accuracy, 0.13 MCC and 0.622 AUC for window length of 25 (Fig. 3, Supplementary Table S1). The CPP approach has been also used in the past, where it was showed that nucleotide/amino acid composition of positive and negative patterns can also use to discriminate these patterns [34,33]. In the MNCPP approach, performance decreased slightly compare to BPP and achieved 53.69% sensitivity, 52.51% specificity, 52.76% accuracy, 0.05 MCC and 0.564 AUC for window length of 21 (Supplementary Fig. S5, Supplementary Table S2). Performance increased significantly when we applied DNCPP and achieved maximum 74.81% sensitivity, 76.72% specificity, 76.31% accuracy, 0.45 MCC and 0.832 AUC for window length of 21 (Fig. 5, Supplementary Table S3). It may be because there were different nucleotides preferred in the positive and negative patterns. We observed that AC, CA, GA and UU di-nucleotide preferred in positive patterns, whereas AA, CG, GC and GG preferred in negative patterns (Fig. 4). Finally, TNCPP achieved highest 83.92% sensitivity, 84.82% specificity, 84.62% accuracy, 0.62 MCC and 0.889 AUC for window length of 19 (Fig. 7, Supplementary Table S4). Here also tri-nucleotide wise preferences were present, where ACA, ACC, AGA, CAC, CCA, GAG, UGA and UUU tri-nucleotides preferred in positive patterns, whereas AAG, AGG, CCG, CGC, GCG, GGC, GGG, GGU, UAA, UGG and UUC preferred in negative patterns (Fig. 6).

Table 1
The prediction performance of different classifiers using TNCPP approach of 19-length window size.

Name of classifier	Sensitivity	Specificity	Accuracy	MCC
NaiveBayes	52.47	56.35	55.50	0.07
NaiveBayesMultinomial	49.44	61.49	58.85	0.09
BayesNet	50.77	60.93	58.71	0.10
ComplementNaiveBayes	41.05	70.01	63.67	0.10
MultilayerPerceptron	54.20	64.57	62.30	0.16
J48	55.30	84.10	77.79	0.38
SMO	39.65	95.96	83.63	0.46
RandomForest	78.03	76.54	76.87	0.47
IBk	76.54	81.99	80.80	0.52
SVM^{light}	83.92	84.82	84.62	0.62

In the present scenario, prediction of protein interacting nucleotides is in the primitive stage. We tried various approaches and achieved reasonable performance, but this problem requires more attention and information, in order to develop an efficient prediction model. The present method has many limitations due to limited dataset availability and criterion to determine protein interacting and non-interacting nucleotides. We used 5.0 Å as a cutoff distance because in the past various methods have been used in this criterion for selecting RNA-interacting residues in proteins [30] but this is not solely a correct criterion. Recently, pseudo k-tuple nucleotide composition (PseKNC) based approach also proposed for different nucleotide related problems [46,47] and our approach can be extended in future works. Additionally, there are sequence-independent bindings also present in the RNA–protein interaction; therefore, it is important to solve these issues in the future. We hope that *RNApin* method will be useful for the RNA biologist in order to identify protein interacting nucleotides in RNA sequences.

5. Conclusion

In this study, we tried various approaches for the prediction of PINs. We optimized different window sizes, SVM parameters and kernels. Finally we found that tri-nucleotide wise compositional differences were present between positive and negative patterns and TNCPP approach was most efficient to discriminate PINs and non-PINs.

RNApin web-server

We implemented SVM prediction model in a web-server called *RNApin*. The RNApin is user-friendly and freely available from http://crdd.osdd.net/raghava/rnapin/ web-address. We have provided our dataset 'RNA-208' in the supplementary file 2 (RNA-208.txt) and also RNA-1546 is accessible from our RNApin webserver (http://crdd.osdd.net/raghava/rnapin/dataset.php).

Author's contributions

BP created dataset, optimized and developed the SVM models. BP also created the backend web server and the front end user interface. GPSR conceived the project, coordinated it and refined the manuscript drafted by BP. Both authors have read and approved the final draft of the manuscript.

Please cite this article as: B. Panwar, G.P.S. Raghava, Genomics (2015), http://dx.doi.org/10.1016/j.ygeno.2015.01.005

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

Authors are thankful to funding agencies Council of Scientific and Industrial Research (project OSDD and GENESIS BSC0121) and Department of Biotechnology (project BTISNET), Govt. of India. This report has Institute of Microbial Technology (IMTECH) communication no. 026/2013.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2015.01.005.

References

[1] D.E. Draper, Themes in RNA–protein recognition, J. Mol. Biol. 293 (1999) 255–270. http://dx.doi.org/10.1006/jmbi.1999.2991.
[2] N. Morozova, J. Allers, J. Myers, Y. Shamoo, Protein–RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures, Bioinformatics 22 (2006) 2746–2752. http://dx.doi.org/10.1093/bioinformatics/btl470.
[3] J. Iwakiri, H. Tateishi, A. Chakraborty, P. Patil, N. Kenmochi, Dissecting the protein–RNA interface: the role of protein surface shapes and RNA secondary structures in protein–RNA recognition, Nucleic Acids Res. 40 (2012) 3299–3306. http://dx.doi.org/10.1093/nar/gkr1225.
[4] A.C. Cheng, W.W. Chen, C.N. Fuhrmann, A.D. Frankel, Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains, J. Mol. Biol. 327 (2003) 781–796 (http://www.ncbi.nlm.nih.gov/pubmed/12654263, accessed February 01, 2013).
[5] A.M. Khalil, J.L. Rinn, RNA–protein interactions in human health and disease, Semin. Cell Dev. Biol. 22 (2011) 359–365. http://dx.doi.org/10.1016/j.semcdb.2011.02.016.
[6] J.C. Darnell, C.E. Fraser, O. Mostovetsky, G. Stefani, T.A. Jones, S.R. Eddy, et al., Kissing complex RNAs mediate interaction between the fragile-X mental retardation protein KH2 domain and brain polyribosomes, Genes Dev. 19 (2005) 903–918. http://dx.doi.org/10.1101/gad.1276805.
[7] F. Bolognani, N.I. Perrone-Bizzozero, RNA–protein interactions and control of mRNA stability in neurons, J. Neurosci. Res. 86 (2008) 481–489. http://dx.doi.org/10.1002/jnr.21473.
[8] S.J. Kolb, D.J. Battle, G. Dreyfuss, Molecular functions of the SMN complex, J. Child Neurol. 22 (2007) 990–994. http://dx.doi.org/10.1177/0883073807305666.
[9] R.N. Kanadia, J. Shin, Y. Yuan, S.G. Beattie, T.M. Wheeler, C.A. Thornton, et al., Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 11748–11753. http://dx.doi.org/10.1073/pnas.0604970103.
[10] C.K. Iwahashi, D.H. Yasui, H.-J. An, C.M. Greco, F. Tassone, K. Nannen, et al., Protein composition of the intranuclear inclusions of FXTAS, Brain 129 (2006) 256–271. http://dx.doi.org/10.1093/brain/awh650.
[11] L.M. Hellman, M.G. Fried, Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions, Nat. Protoc. 2 (2007) 1849–1861. http://dx.doi.org/10.1038/nprot.2007.249.
[12] J. Ule, K. Jensen, A. Mele, R.B. Darnell, CLIP: a method for identifying protein–RNA interaction sites in living cells, Methods 37 (2005) 376–386. http://dx.doi.org/10.1016/j.ymeth.2005.07.018.
[13] A. Günzl, Z. Palfi, A. Bindereif, Analysis of RNA–protein complexes by oligonucleotide-targeted RNase H digestion, Methods 26 (2002) 162–169. http://dx.doi.org/10.1016/S1046-2023(02)00019-1.
[14] J.D. Keene, J.M. Komisarow, M.B. Friedersdorf, RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts, Nat. Protoc. 1 (2006) 302–307. http://dx.doi.org/10.1038/nprot.2006.47.
[15] S.A. Tenenbaum, P.J. Lager, C.C. Carson, J.D. Keene, Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays, Methods 26 (2002) 191–198. http://dx.doi.org/10.1016/S1046-2023(02)00022-1.
[16] D.L. Beach, J.D. Keene, Ribotrap: targeted purification of RNA-specific RNPs from cell lysates through immunoaffinity precipitation to identify regulatory proteins and RNAs, Methods Mol. Biol. 419 (2008) 69–91. http://dx.doi.org/10.1007/978-1-59745-033-1_5.
[17] M. Kumar, M.M. Gromiha, G.P.S. Raghava, Prediction of RNA binding sites in a protein using SVM and PSSM profile, Proteins 71 (2008) 189–194. http://dx.doi.org/10.1002/prot.21677.
[18] M. Terribilini, J.D. Sander, J.-H. Lee, P. Zaback, R.L. Jernigan, V. Honavar, et al., RNABindR: a server for analyzing and predicting RNA-binding sites in proteins, Nucleic Acids Res. 35 (2007) W578–W584. http://dx.doi.org/10.1093/nar/gkm294.
[19] L. Wang, C. Huang, M.Q. Yang, J.Y. Yang, BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, BMC Syst. Biol. 4 (Suppl. 1) (2010) S3. http://dx.doi.org/10.1186/1752-0509-4-S1-S3.
[20] R.V. Spriggs, Y. Murakami, H. Nakamura, S. Jones, Protein function annotation from sequence: prediction of residues interacting with RNA, Bioinformatics 25 (2009) 1492–1497. http://dx.doi.org/10.1093/bioinformatics/btp257.
[21] S. Choi, K. Han, Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets, Comput. Biol. Med. 43 (2013) 1687–1697. http://dx.doi.org/10.1016/j.compbiomed.2013.08.011.
[22] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, et al., iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529. http://dx.doi.org/10.1093/bioinformatics/btu083.
[23] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, K.-C. Chou, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo tri-nucleotide composition, Anal. Biochem. 462 (2014) 76–83. http://dx.doi.org/10.1016/j.ab.2014.06.022.
[24] W.-R. Qiu, X. Xiao, K.-C. Chou, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766. http://dx.doi.org/10.3390/ijms15021746.
[25] K.-C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247. http://dx.doi.org/10.1016/j.jtbi.2010.12.024.
[26] H. Lin, E.-Z. Deng, H. Ding, W. Chen, K.-C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972. http://dx.doi.org/10.1093/nar/gku1019.
[27] B. Liu, F. Liu, L. Fang, X. Wang, K.-C. Chou, repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, Bioinformatics (2014) http://dx.doi.org/10.1093/bioinformatics/btu820.
[28] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, K.-C. Chou, PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (2014) 119–120. http://dx.doi.org/10.1093/bioinformatics/btu602.
[29] B.A. Lewis, R.R. Walia, M. Terribilini, J. Ferguson, C. Zheng, V. Honavar, et al., PRIDB: a protein–RNA interface database, Nucleic Acids Res. 39 (2011) D277–D282. http://dx.doi.org/10.1093/nar/gkq1108.
[30] R.R. Walia, C. Caragea, B.A. Lewis, F. Towfic, M. Terribilini, Y. El-Manzalawy, et al., Protein–RNA interface residue prediction using machine learning: an assessment of the state of the art, BMC Bioinf. 13 (2012) 89. http://dx.doi.org/10.1186/1471-2105-13-89.
[31] F. Ahmed, G.P.S. Raghava, Designing of highly effective complementary and mismatch siRNAs for silencing a gene, PLoS ONE 6 (2011) e23443. http://dx.doi.org/10.1371/journal.pone.0023443.
[32] B. Panwar, G.P. Raghava, Prediction of uridine modifications in tRNA sequences, BMC Bioinf 15 (2014) 326. http://dx.doi.org/10.1186/1471-2105-15-326.
[33] H.R. Ansari, G.P. Raghava, Identification of conformational B-cell epitopes in an antigen from its primary sequence, Immunome Res. 6 (2010) 6. http://dx.doi.org/10.1186/1745-7580-6-6.
[34] S. Agarwal, N.K. Mishra, H. Singh, G.P.S. Raghava, Identification of mannose interacting residues using local composition, PLoS ONE 6 (2011) e24039. http://dx.doi.org/10.1371/journal.pone.0024039.
[35] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (1999) 988–999. http://dx.doi.org/10.1109/72.788640.
[36] T. Joachims, Making large-scale SVM learning practical, Adv. Kernel Methods Support Vector Learn1999. 169–184. http://dx.doi.org/10.1109/ICEMI.2009.5274151.
[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. 11 (2009) 10–18. http://dx.doi.org/10.1145/1656274.1656278.
[38] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349. http://dx.doi.org/10.3109/10409239509083488.
[39] K.-C. Chou, H.-B. Shen, Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162. http://dx.doi.org/10.1038/nprot.2007.494.
[40] B. Panwar, S. Gupta, G.P. Raghava, Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information, BMC Bionf. 14 (2013) 44. http://dx.doi.org/10.1186/1471-2105-14-44.
[41] B. Panwar, G.P.S. Raghava, Prediction and classification of aminoacyl tRNA synthetases using PROSITE domains, BMC Genomics 11 (2010) 507. http://dx.doi.org/10.1186/1471-2164-11-507.
[42] B. Panwar, A. Arora, G.P.S. Raghava, Prediction and classification of ncRNAs using structural information, BMC Genomics 15 (2014) 127. http://dx.doi.org/10.1186/1471-2164-15-127.
[43] M.M. Gromiha, K. Yokota, K. Fukui, Understanding the recognition mechanism of protein–RNA complexes using energy based approach, Curr. Protein Pept. Sci. 11 (2010) 629–638 (http://www.ncbi.nlm.nih.gov/pubmed/20887255, accessed November 30, 2014).
[44] J.S. Chauhan, N.K. Mishra, G.P.S. Raghava, Identification of ATP binding residues of a protein from its primary sequence, BMC Bioinf 10 (2009) 434. http://dx.doi.org/10.1186/1471-2105-10-434.
[45] B. Panwar, G.P.S. Raghava, Predicting sub-cellular localization of tRNA synthetases from their primary structures, Amino Acids 42 (2012) 1703–1713. http://dx.doi.org/10.1007/s00726-011-0872-8.
[46] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, K.-C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60. http://dx.doi.org/10.1016/j.ab.2014.04.001.
[47] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68. http://dx.doi.org/10.1093/nar/gks1450.