JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 1, 2017
Mary Ann Liebert, Inc.
Pp. 31–39
DOI: 10.1089/cmb.2016.0128

A Graph Approach to Mining Biological Patterns in the Binding Interfaces

WEN CHENG and CHANGHUI YAN

ABSTRACT

Protein–RNA interactions play important roles in the biological systems. Searching for regular patterns in the Protein–RNA binding interfaces is important for understanding how protein and RNA recognize each other and bind to form a complex. Herein, we present a graph-mining method for discovering biological patterns in the protein–RNA interfaces. We represented known protein–RNA interfaces using graphs and then discovered graph patterns enriched in the interfaces. Comparison of the discovered graph patterns with UniProt annotations showed that the graph patterns had a significant overlap with residue sites that had been proven crucial for the RNA binding by experimental methods. Using 200 patterns as input features, a support vector machine method was able to classify protein surface patches into RNA-binding sites and non-RNA-binding sites with 84.0% accuracy and 88.9% precision. We built a simple scoring function that calculated the total number of the graph patterns that occurred in a protein–RNA interface. That scoring function was able to discriminate near-native protein–RNA complexes from docking decoys with a performance comparable with that of a state-of-the-art complex scoring function. Our work also revealed possible patterns that might be important for binding affinity.

Keywords: binding sites, common subgraphs, graph patterns, protein–RNA interactions, recurrent patterns, scoring functions.

1. INTRODUCTION

It is well known that protein–RNA interactions play important roles in various biological processes. Understanding the molecular foundation of the interactions will give us the ability to modify the interfaces to regulate those biological processes. This study aims to analyze protein–RNA interfaces to identify recurrent structural patterns facilitating the interactions. Various proteomic projects have produced a large number of protein structures, whose functions are still unknown. The patterns discovered in this study will be helpful for discovering RNA-binding sites on those protein structures and predicting the structures of the protein–RNA complexes that they may form.

Many computational methods have been developed for predicting RNA-binding sites on proteins. Some of them used the machine-learning approach to train various computational methods to build classifiers

Department of Computer Science, North Dakota State University, Fargo, North Dakota.

that could predict RNA-binding sites on the protein structures using sequence, structural, and evolutionary features as input (Terribilini et al., 2006, 2007; Liu et al., 2010; Murakami et al., 2010; Chen et al., 2014; Yang et al., 2014). Although the classifiers reportedly achieved accurate predictions in many cases, a limitation of these methods is that the predicting process is like a black box, and it is hard to translate the rules used by the classifiers into knowledge to elucidate the affinity and specificity of the interactions. Other methods used structural alignment to transfer known RNA-binding sites from a homologous template structure to a structure of interest. This approach requires an available homologous template, whose RNA-binding sites are known. However, in many cases, this requirement cannot be satisfied.

Another important problem in the study of protein–RNA interactions is to predict the conformation of the complex that a protein and RNA may form in the interaction. Docking is a popular approach to solving this problem. The docking approach generates a large set of poses that represent the whole universe of possible conformations of the protein–RNA complex and then uses a scoring function to rank the poses. A good scoring function should assign higher scores to poses similar to the native structure and lower scores to poses that are dissimilar. Various scoring functions have been used. One important category of scoring functions is called knowledge-based scoring function, which reflects the propensities of a certain moieties to occur in the protein–RNA interfaces in a database of known protein–RNA complexes. The moieties could be atoms, atom pairs, residues, residue pairs, or others (Pérez-Cano and Fernández-Recio, 2010; Zhao et al., 2011; Li et al., 2012; Huang and Zou, 2014).

In this work, we used a graph-mining method to discover graph patterns in the protein–RNA interfaces. Our results showed that the graph patterns covered residue sites that had been proven crucial for the interactions by experimental methods. We demonstrated that the discovered patterns could be used to predict RNA-binding sites on protein structures with high accuracy and high precision, and they could also be used as a scoring function to discriminate near-native protein–RNA complexes from docking decoys. Our work also revealed possible patterns that might be important for binding affinity.

We have significantly extended the work beyond previous work (Cheng and Yan, 2015). First, we explored different machine learning methods to build classifiers for the discrimination of RNA-binding sites versus non-RNA-binding sites. Second, we compared the discovered graph patterns with the UniProt (Bairoch et al., 2005) MUTAGEN annotations, which were collected from mutagenesis experiments and contained information about how alteration of a residue would affect the binding between protein and RNA. Compared with the UniProt REGION annotations, the MUTAGEN annotations have higher precision. Third, we investigated how the performance of the proposed scoring function varies, as different numbers of patterns were used in the scoring function. This led to an important discovery (see Subsection 3.4 of Section 3) that some patterns were more important for the scoring function, which suggested that these patterns might contribute more to the binding affinity. Finally, we attempted to improve the scoring function using weighted sum of the patterns. Our analysis suggested that the order of the patterns that were sorted based on enrichment levels did not strictly reflect the order of importance of the patterns' contribution to the binding affinity.

2. METHODS

2.1. Data sets

Our study used two data sets. The first data set, will be referred to as Data set I, included a set of three-dimensional structures of protein–RNA complexes that had been determined using experimental methods. Each complex structure showed a native binding mode between a protein and an RNA. In this study, the training set was used to discover common subgraphs enriched in the RNA-binding sites. Data set I was obtained from the RCSB Protein Data Bank (PDB) database. First, we retrieved from PDB representatives of protein–RNA complexes with no more than 90% sequence identity. The search returned 1570 hits. Then, the data set was culled using PISCES (Wang and Dunbrack, 2003) with the mutual sequence similarity no more than 25%, maximum resolution of 3 Å, maximum R-value of 0.3, minimum length of 40 amino acids, and maximum length of 1000. After the culling, 144 protein–RNA complexes remained. The second data set, will be referred to as Data set II, was derived from a protein–RNA docking benchmark collected by Huang and Zou (2013). The original data set from Huang and Zou (2013) was a nonredundant set of 72 protein–RNA complexes and their unbound structures. In this study, we removed the proteins that overlapped with Data set I. At the end, Data set II consisted of 37 protein–RNA complexes and their unbound structures.

2.2. Interface residues

Interface residues on the RNA-binding sites were defined as in Jones et al. (2003). We used NACCESS software to calculate the accessible surface area (ASA) of each amino acid in both bounded and unbounded states. An amino acid was defined as an interface residue if its ASA in unbounded state was at least 1 $Å^2$ more than that in bounded state.

2.3. Graph representation of RNA-binding sites

Each RNA-binding site was represented using a graph, where each node represented an interface residue, and an edge was added between two nodes if the corresponding residues were in contact. Two residues were considered contacting if the nearest distance between their heavy atoms was less than 0.5 Å plus the atoms' radii. Each node was labeled with its residue type. Each edge was also associated with an edge label. If the two nodes at the end of an edge were sequence neighbors on the protein chain, then the edge was labeled as type one; otherwise, the edge was labeled as type two.

2.4. Discovery of common subgraphs

There were 144 RNA-binding sites in Data set I, and each one was represented as a graph. We will refer to these graphs as binding-site graphs. We implemented the VF2 algorithm (Cordella et al., 2004) to find common subgraphs between each pair of binding-site graphs. In the test of isomorphism, we also took into consideration the node labels and edge labels. In this study, we focused on the common subgraphs of sizes 3 and 4, that is, each common subgraph had three nodes or four nodes, as common subgraphs with less than three nodes contain too few information and common subgraphs with more than five nodes were rarely found.

2.5. Classification

We built machine-learning classifiers to distinguish RNA-binding sites from non-RNA-binding sites. The classification was evaluated using fivefold cross validation at protein level. The whole data set was split into five subsets at the protein levels, such that surface patches from the same protein remained in the same subset. In each round of experiment, one subset was used as test set and the other four as training set. Five rounds of experiments were conducted so that each subset was used as test set once. We tried different machine-learning algorithms, including the decision tree (J48) (Quinlan, 1986) and Random Forest (Breiman, 2001), implemented in Weka (Hall et al., 2009), and LibSVM (Chang and Lin, 2011).

The classification performance was evaluated using the following metrics: True Positive (TP): RNA-binding sites that were predicted as RNA-binding sites; False Positive (FP): nonbinding sites that were predicted as RNA-binding sites; False Negative (FN): RNA-binding sites that were predicted as nonbinding sites; True Negative (TN): nonbinding sites that were predicted as nonbinding sites; Accuracy: (TP+TN)/(TP+TN+FP+FN); Precision: TP/(TP+FP); and AUC: Area under the ROC curve.

3. EXPERIMENTS AND RESULTS

3.1. Discovery of graph patterns enriched in RNA-binding sites

Our goal was to discover graph patterns enriched in the RNA-binding sites, that is, graph patterns that occurred with high frequencies in RNA-binding sites and with low frequencies in the rest of the protein surface. We found common subgraphs in Data set I, as described in Section 2.4 of Section 2. After removing duplicated common subgraphs, we obtained 3363 unique subgraphs of size 3 and 7482 unique subgraphs of size 4. These subgraphs represented some graph patterns observed in the RNA-binding sites. We randomly collected 144 nonbinding sites from the 144 proteins, with one nonbinding site from each protein. The nonbinding site from a protein had the same size as the RNA-binding site from the same protein, and there was no overlap between the nonbinding site and RNA-binding site. These nonbinding sites served as the background for the identification of patterns enriched in the RNA-binding sites.

For each subgraph, we checked whether it occurred in the 144 binding-site graphs and the 144 nonbinding sites. The presence or absence of a subgraph in the RNA-binding-site and nonbinding sites was recorded using a vector of 288 values, with 1 being presence and 0 absence. Then, we performed a t-test to identify subgraphs that enriched in the RNA-binding sites. A lower p-value given by the t-test indicated that

Table 1. Classification of RNA-Binding Sites Versus Nonbinding Sites Using libSVM

Size of patterns	*No. of patterns*	*TP*	*FP*	*FN*	*TN*	*Accuracy (%)*	*Precision (%)*	*AUC*
Three nodes	Top 100	108	48	36	96	74.0	69.2	0.71
	Top 200	94	28	50	116	77.4	77.0	0.73
	Top 300	101	30	43	114	76.7	77.1	0.75
	Top 400	93	13	51	131	77.8	87.8	0.78
	Top 500	95	16	49	128	76.4	85.6	0.77
Four nodes	Top 100	88	15	56	129	76.0	85.4	0.75
	Top 200	112	14	32	130	84.0	88.9	0.84
	Top 300	118	29	26	115	80.9	80.3	0.81
	Top 400	109	21	35	123	80.9	83.8	0.81

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

the subgraph was more favored by the RNA-binding sites. We sorted the subgraphs in the order of increasing *p*-values. Thus, the subgraphs at the top of the list were more favored at the RNA-binding sites.

3.2. Predicting RNA-binding sites using the enriched graph patterns

In the above section, we have discovered a list of graph patterns with various enrichment levels in the RNA-binding sites. To further evaluate the significance of the graph patterns, we tested the feasibility of using the graph patterns to discover RNA-binding sites on the protein surface. For this purpose, we used the graph patterns as input to train a support vector machine classifier to discriminate RNA-binding sites from nonbinding sites. When *n* patterns were chosen, an RNA-binding site or nonbinding site was encoded as a vector of *n* values, representing the presence or absence of the first *n* patterns from the top of the list. The classification was evaluated using fivefold cross validation at protein level. libSVM with radial basis function (RBF) kernel was used. We tried different numbers of patterns from 100 to 500 with an increment of 100 for three-node subgraphs and four-node subgraphs separately. The classification performance is shown in Table 1. Our results showed that as the number of patterns increased, the accuracy first increased and then decreased. Using four-node subgraphs as input, the classifier was able to achieve better performance than using three-node subgraphs. When the first 200 four-node subgraphs from the list were used, the classifier achieved the best performance with 84.0% accuracy, 88.9% precision, and 0.84 AUC. These results suggested that the enriched subgraphs we discovered revealed structural patterns that facilitated the interactions between protein and RNA and thus could be used to predict RNA-binding sites.

We also used Random Forest and decision tree methods to build the classifiers. Table 2 shows the comparison between these algorithms when 200 four-node subgraphs were used to encode surface sites. The Random Forest achieved slightly higher precision than libSVM, but with much lower accuracy. The decision tree had the lowest accuracy and precision among the three algorithms. When different numbers of subgraphs were used, the same trend was observed when libSVM was used.

3.3. Significant overlap between graph patterns and UniProt annotations

To further evaluate the biological significance of the discovered graph patterns, we compared them with the annotations in UniProt, a comprehensive database of protein functional information. The REGION annotation in the UniProt denoted the stretch of protein sequence involved in a certain type of function. We focused on the REGION annotations that were associated with the RNA-binding function and compared the residues covered by the graph patterns and the residues covered by the RNA-binding REGIONs. For

Table 2. Comparison Between Different Classification Algorithms

Algorithms	*TP*	*FP*	*FN*	*TN*	*Accuracy (%)*	*Precision (%)*	*AUC*
Decision tree (J48)	115	116	29	28	49.6	49.7	0.50
Random Forest	73	7	71	137	72.9	91.2	0.73
libSVM	112	14	32	130	84.0	88.9	0.84

Two hundred 4-node subgraphs were used to encode surface sites.

Table 3. Overlap Between Subgraph Patterns and Uniprot Regions

PDBID	*No. of residues in subgraphs*	*No. of residues in REGION annotations*	*No. of residues overlapped*
1YVP	14	165	12
1K8W	17	29	4
1KNZ	4	146	4
3MOJ	5	76	5
4IG8	13	59	8
2ZKO	10	73	10
1H4S	4	30	0
3DH3	22	8	4
2A8V	4	17	1
3FOZ	19	28	13
3RW6	13	117	0
4KXT	20	95	14
1JID	4	9	0
2BH2	16	24	6
1N78	16	16	0
3MDI	14	146	10

simplicity, we only looked at the top 200 four-node graph patterns, since the previous section showed that these patterns gave the best performance in classification. Among the 144 proteins in Data set I, 16 have a REGION annotation associated with RNA binding and also include at least one pattern from the top 200 four-node patterns. The number of residues covered by the patterns, the number of residues covered by the REGION, and the overlap between the two sets are shown in Table 3. The table showed that in 12 of the 16 proteins, the graph patterns overlap with the REGIONs. For the 16 proteins, the average size of the proteins was 319.5, and the average sizes of the REGIONs and patterns were 64.8 and 12.1, respectively, and the average overlap was 5.6. The overlap had a p-value of 0.02, that is, if the REGIONs and patterns were randomly generated, then there was only a probability of 0.02 to achieve overlap equal to or better than this. This result strongly supports the biological significance of the discovered patterns.

The REGION annotation usually contains a contiguous segment on the protein sequence that is believed to be associated with a function. They often include many residues that do not directly participate in the RNA-binding function. Thus, it is understandable that the REGION annotations cover much more residues than the subgraph patterns.

The UniProt database also includes MUTAGEN annotations and describes how experimental mutations of one or more amino acids change the biological properties of the protein. In our data set, we searched for the MUTAGEN annotations, whose mutation reduced the RNA-binding ability of the protein. Different than the REGION that includes a long stretch of residue, MUTAGEN only includes a few isolated residues. Among the 144 proteins in Data set I, 21 have at least one MUTAGEN annotation, associated with RNA binding, and also include at least one pattern from the top 200 four-node patterns. In eight of them, the MUTAGEN and the subgraph overlapped (Table 4). This overlap had a p-value of 0.005. This confirms the

Table 4. Overlap Between Subgraph Patterns and Uniprot Mutagens

PDBID	*No. of residues in subgraphs*	*No. of residues in MUTAGEN annotations*	*No. of residues overlapped*
2F8K	5	2	1
1FEU	9	8	4
2XS2	8	4	3
2A1R	4	4	1
2A8V	4	2	2
2BGG	10	5	3
3PEY	4	8	2
3MDI	14	5	2

biological significance of the subgraph patterns. We also noticed that in 13 proteins, there was no overlap between MUTAGEN annotations and the subgraph patterns. One possible explanation for this is that the information gathered from mutagenesis experiments is scarce, and thus, the MUTAGEN annotations only cover a small fraction of RNA-binding residues. Meanwhile, the subgraph patterns were meant to capture patterns that were important for the binding, and thus, they also only cover a small fraction of the RNA-binding residues. Thus, there is a high chance that the two sets may not overlap.

3.4. Discrimination of near-native protein–RNA conformations from docking decoys

Molecular docking is a very popular approach to predicting the structure of complex that a protein and an RNA may form when their unbound structures are available. In the first step of the docking, a large number of poses are generated, with each pose representing one possible conformation the protein–RNA complex may take. In the second step, a scoring function is used to assign scores to the poses. The poses that are very similar to the native protein–RNA complexes are usually called near-native poses, and the poses that are not similar to the native complexes are called docking decoys. A good scoring function should assign higher scores to near-native poses than to the docking decoys.

In previous sections, we have discovered a list of graph patterns that are favored in the protein–RNA interfaces. In this study, we will test these patterns' ability to discriminate near-native structures from docking decoys. We built a very simple scoring function based on the occurrences of the patterns in the protein–RNA interfaces. For a docking pose, the proposed scoring function counted the number of these graph patterns that occurred on the protein–RNA interface and assigned a score equal to the number. We evaluated the scoring function using Data set II, a docking benchmark collected by Huang and Zou (2013), and compared it with the scoring function used in 3dRPC, a state-of-the-art protein–RNA docking method developed by Huang et al. (2013). They compared 3dRPC with other docking methods using Data set II, and the results showed that 3dRPC was better than the others.

3dRPC consisted of two parts: RNA-Protein (RP)-Dock, which generated potential poses for the protein–RNA complexes, and a distance- and environment-dependent, coarse-grained, and knowledge-based potential for RNA-Protein (DECK-RP). Huang et al. (2013) generated 1000 poses using RP-Dock for each protein and RNA pair. Each pose was aligned with the native protein–RNA structure by superimposing the protein, and if the root mean square deviation between the pose and native structure is less than 10 Å, then the pose was considered a near-native pose. Then, they used the DECK-RP to rank the 1000 poses and predicted Np best poses to be near-native poses, where Np will be referred to as the prediction number. Each protein and RNA pair in Data set II were considered as one test case. They evaluated the performance using success rate and hit count, where success rate was the fraction of the test cases where the top Np poses contained at least one near-native pose, and hit number was the mean number of near-native poses within the top Np poses calculated as the average over all test cases.

We followed the evaluation procedure used by Huang et al. (2013). We used the same poses generated by them and used our simple scoring method to rank them. There is still one parameter in our scoring function that needs to be decided. We have a list of graph patterns sorted by enrichment, but we have not decided how many graph patterns the scoring function should consider. For simplicity, we only used three-node patterns for this test. To decide this parameter, we set prediction number Np to be 100 and tested how success rate and hit count changed when we increased the number of graph patterns that the scoring function considered. We started from 100 patterns with an increment of 100. Figure 1 showed that at the beginning, both success rate and hit number increased slowly as the pattern number increased, and when the number of patterns increased to 1200, both success rate and hit number reached the highest levels and only fluctuated slightly when the pattern number continued to increase. Thus, in the following comparison, our scoring function considered the first 1200 graph patterns from the list. One interesting observation in Figure 1 is that when pattern number increased from 700 to 800, both success rate and hit number increased dramatically. We tried different prediction number Np from 10 to 900. The same trend was observed. This may suggest that patterns falling in the range from 700 to 800 on the sorted list are crucial for achieving strong binding affinity between proteins and RNA, that is, they contribute more to the binding affinity. Further investigations are needed to reveal how these patterns contribute to the binding affinity.

Figure 2 compares the success rate and hit count between our scoring method and DECK-RP over a wide range of prediction number, from 1 to 900. The results show that the performance of our method is comparable to that of DECK-RP over the whole range. Given that our scoring method only keeps a tally

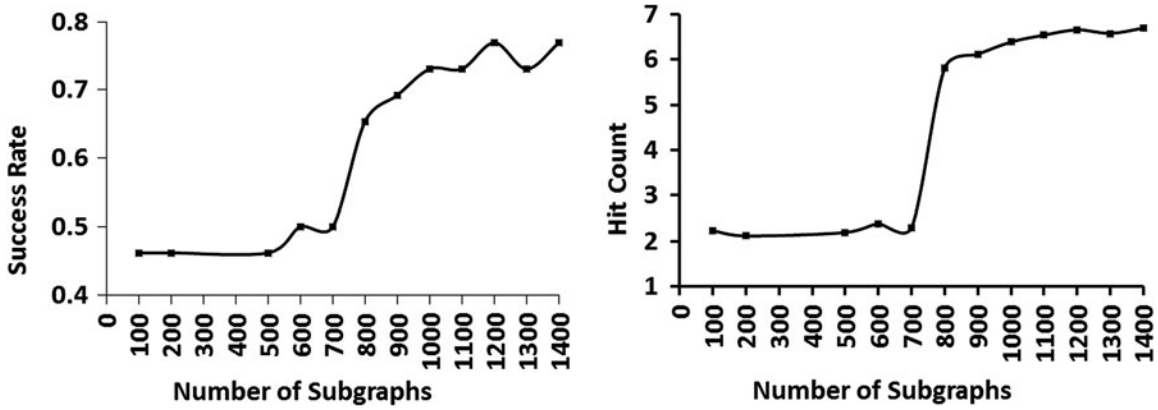

FIG. 1. Success rate and hit count varied when different numbers of subgraph patterns were used in our scoring method. Both measurements reached the highest levels when 1200 patterns were used. There was a hike in both measurements when the pattern number changed from 700 to 800.

count of graph patterns and DECK-RP has to calculate a complicated energy function that includes many parameters, it is exciting to see that our simple method is comparable to DECK-RP. This also suggests the importance of the discovered patterns to the protein–RNA interactions.

In an effort to improve the performance of the scoring method, we assigned the patterns different weights, according to their enrichment levels, in the RNA-binding interfaces, so that patterns with higher enrichment levels at the protein–RNA interfaces have higher weights. Then, the scoring function was calculated as a weighted sum of the patterns. However, this approach slightly decreased the success rate and hit count. One possible explanation is that the patterns that occur most frequently at the interfaces may not contribute the most to the binding affinity. This speculation is consistent with the observation in Figure 1 that patterns falling in the range from 700 to 800 seem to be most useful for scoring the docking poses. It is worth noting that Table 1 shows that the top 200 patterns were most important for discriminating RNA-binding sites from nonbinding sites in the classification test. This result from Table 1 does not contradict with the speculation we make here. Because the classification methods only exploit patterns' distribution in the two classes, the patterns that show bigger difference in their enrichment levels in the two classes are more important for the classification task. The classification does not consider the contribution of each pattern to the binding affinity. Thus, the list of subgraph patterns sorted using *t*-test reflects the patterns' propensities to occur in the RNA-binding sites, and the ranking of the patterns is useful for discriminating binding sites from nonbinding sites. However, the order of the list does not necessarily reflect the order of importance of these patterns' contribution to the binding affinity. Further investigations are needed to determine how these patterns contribute to the binding affinity and, possibly, binding specificity.

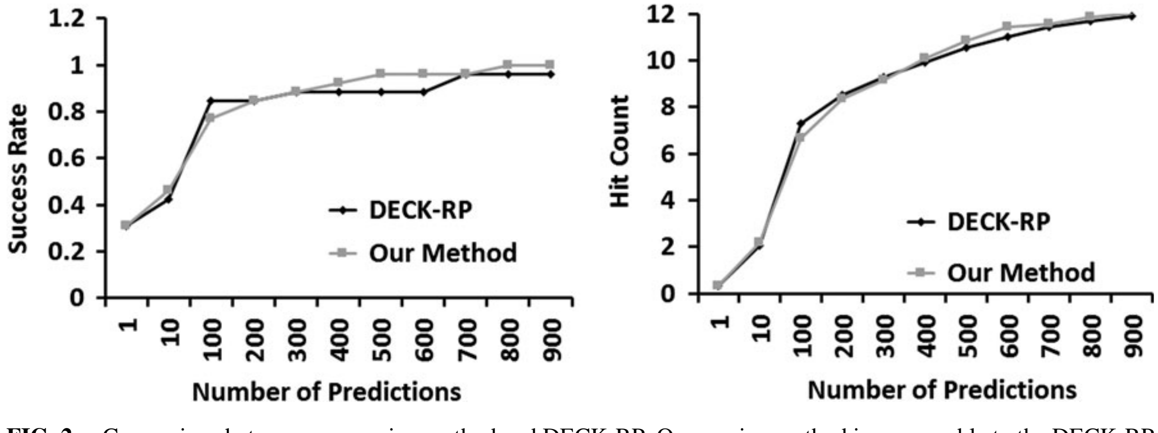

FIG. 2. Comparison between our scoring method and DECK-RP. Our scoring method is comparable to the DECK-RP method in the whole range of predictions numbers. DECK-RP, .

4. CONCLUSIONS

In this work, we discovered graph patterns that were enriched in the protein–RNA interfaces. These patterns were favored in the protein–RNA interfaces and were depleted at the rest of the protein surface. We validated the importance of these patterns in three experiments. In the first, we showed that these patterns could be used to predict RNA-binding sites with high accuracy (84.0%) and precision (88.9%). In the second, we showed that the patterns had a significant overlap with known RNA-binding residues as annotated in UniProt. The third experiment showed that the patterns could be used to discriminate near-native docking poses from docking decoys with performance comparable to a state-of-the art method. Our work also revealed possible patterns that might be important for achieving binding affinity. Our method for pattern mining and the patterns discovered in this study will be very useful for the investigation of interaction mechanisms between protein and RNA and other macromolecules.

ACKNOWLEDGMENTS

This work is sponsored by NSF grants 1262265 and 1305175 to C.Y. C.Y. conceived the project, designed the experiments, and led data analysis and manuscript writing. W.C. carried out all programming and participated in data analysis, discussions, and drafting the manuscript. All authors approved the final version of this manuscript.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bairoch, A., Apweiler, R., Wu, C.H., et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33, D154–D159.

Breiman, L. 2001. Random forests. *Machine Learn*. 45, 5–32.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27.

Chen, Y.C., Sargsyan, K., Wright, J.D., et al. 2014. Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Res*. 42, e15.

Cheng, W., and Yan, C.2015. Mining graph patterns in the protein-RNA interfaces. In IEEE International Conference on Bioinformatics and Biomedicine, Washington, DC, pp. 1267–1271.

Cordella, L., Foggia, P., Sansone, C., et al. 2004. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell*. 26, 1367–1372.

Hall, M., Frank, E., Holmes, G., et al. 2009. The WEKA Data Mining Software: An update. *SIGKDD Explor.* 11, 10–18.

Huang, S.-Y., and Zou, X. 2013. A nonredundant structure dataset for benchmarking protein-RNA computational docking. *J. Comput. Chem.* 34, 311–318.

Huang, S.-Y., and Zou, X. 2014. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res.* 42, e55.

Huang, Y., Liu, S., Guo, D., et al. 2013. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Sci. Rep.* 3, 1887.

Jones, S., Shanahan, H.P., Berman, H.M., et al. 2003. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* 31, 7189–7198.

Li, C.H., Cao, L.B., Su, J.G., et al. 2012. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins* 80, 14–24.

Liu, Z.-P., Wu, L.-Y., Wang, Y., et al. 2010. Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics* 26, 1616–1622.

Murakami, Y., Spriggs, R.V., Nakamura, H., et al. 2010. PiRaNhA: A server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res.* 38, W412–W416.

Pérez-Cano, L., and Fernández-Recio, J. 2010. Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins* 78, 25–35.

Quinlan, R. 1986. Induction of decision tree. *Machine Learn.* 1, 81–106.

Terribilini, M., Lee, J.H., Yan, C., et al. 2006. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* 12, 1450–1462.

Terribilini, M., Sander, J.D., Lee, J.-H., et al. 2007. RNABindR: A server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.* 35, W578–W584.

Wang, G., and Dunbrack, R.L.J. 2003. PISCES: A protein sequence culling server. *Bioinformatics* 19, 1589–1591.

Yang, X.-X., Deng, Z.-L., and Liu, R.2014. RBRDetector: Improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. *Proteins* 82, 2455–2471.

Zhao, H., Yang, Y., and Zhou, Y.2011. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.* 39, 3017–3025.

Address correspondence to:
Dr. Changhui Yan
Department of Computer Science
North Dakota State University
Fargo, ND 58106

E-mail: changhui.yan@ndsu.edu