Systematic Identification of Local Structure Binding Motifs in Protein-RNA Recognition

Zhi-Ping Liu*
*Department of Biomedical Engineering, School of Control Science and Engineering,
Shandong University, Jinan, Shandong 250061, China
Email: zpliu@sdu.edu.cn

***Abstract*—Many critical biological processes are strongly related to protein-RNA interactions. Revealing the structure motifs of performing protein-RNA binding function will provide valuable information for deciphering their interaction mechanisms and benefit complementary structure designs in bioengineering. In this work, we provide a study of systematic identification of protein structure motifs of RNA-binding sites in form of pockets on protein surfaces by clustering these local structure patterns into similar groups. We also identify the crucial recognition patterns and the structural complementary features in the protein-RNA binding events.**

I. Introduction

Protein-RNA recognitions provide crucial interactions of biochemical reactions and signaling transductions in many fundamental biological processes [1], [2]. Revealing the structure motifs of binding events will provide clues for deciphering the mechanisms of protein-RNA interaction and provide valuable knowledge for drug design and protein engineering, such as drug targets of silencing some specific RNAs after transcription [3]. Some sequence patterns have been identified for protein-RNA recognition, such as RNA-recognition motifs and zinc fingers [4]. Some critical partners in the RNA interference of miRNA-binding functions, e.g., Dicer and Argonaute (AGO) [5], also indicate their functional importance and specificity for protein-RNA interaction.

Protein surfaces are known as one of the major places where the RNA-binding events take place [6]. Pockets are one of the significant structural patterns on protein surfaces, which are believed to provide concrete spot and detailed environment for many critical biochemical reactions [7], [8]. Protein binding pocket provides the local structure for packing RNA and constructing a complex with certain functions [9]. Recent studies have made substantial efforts in predicting protein-RNA binding sites, such as PRNA [10], RNABindR [11], BindN [12], and PRINTR [13] , and protein-RNA interactions [14], [15], while few analyses have been implemented to identify the binding features underlying RNA-binding pockets [2], [9], as well as major local structure groups and structure motifs of RNA binding. The knowledge of RNA binding pockets will reveal the RNA binding specificity and mechanism in the protein-RNA recognition. Identification of the physicochemical and structural features of these binding pockets will highly benefit the research of protein-RNA interactions.

In this work, we provide a large-scale analysis of the RNA-binding pockets in proteins for identifying the structural motifs and features in protein-RNA recognition. We firstly identify the RNA-binding sites on protein surfaces and extract the surface cavities involved in the binding events from our compiled non-redundant protein-RNA complexes. The local structure similarities in the RNA-binding pockets and the global structure similarities in their associated proteins are measured by structural alignment algorithms respectively. The sequence and composition domain features are identified at the same time. Clustering in the RNA-binding proteins, domains and pockets has been implemented subsequently to reveal the protein groups, the important binding structure patterns and motifs, and their functional implications in protein-RNA interactions.

II. Results and discussion

A. RNA-binding events mainly take place in pockets on protein surfaces

In the collected non-redundant 158 protein-RNA complexes (see Materials and methods), there are 48484 residues in the 20 types of amino acids, among which 3360 residues are identified as RNA-binding residues which contact the RNA nucleotides detected in the crystallized complexes. Totally, we extract 7664 pockets on these protein surfaces by CASTp [16]. These pockets totally contain 34119 residues, i.e., 70.4% of the total residues, which form the three-dimensional local structure cavities on protein surfaces. A pocket is classified as a RNA-binding pocket when it contains at least one RNA-binding residue. There are 1539 pockets (20% of all the pockets) are involved in the RNA-binding events, which contain 2849 RNA-binding residues (85% of all the RNA-binding residues). There are 16303 amino acid residues (48% of all the residues) are involved in the RNA-binding pockets. The composition percentages demonstrate that the RNA-binding events mainly take place in the pockets on protein surface. The local cavities forming pocket-like shapes provide the structural environment of protein-RNA interactions. To identify RNA-binding structure motifs, the importance of pockets in the RNA binding directs us to focus on these local structures of pockets.

The RNA-binding pockets contain the average of 10 residues which relatively bigger than the average of 6 residues in all the pockets, though the number of amino acid residues contained in a RNA-binding pocket is very diverse in the range of 1 to 528. For omitting the tiny pockets containing fewer than 4 residues, we obtain 786 pockets at least containing 5 amino acid residues. We will focus on our analysis in these pockets. For the 20 types of amino acids, Figure 1 displays their composition ratios in different residue sets. Certain amino acids are found to be favor in the RNA-binding events, such as ‘R’ (Arginine) and ‘K’ (Lysine) are the most preferred

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

amino acids contacting RNA. Interestingly, 'E' (Glutamic acid) and 'L' (Leucine) are not obviously working as RNA-binding residues [17], but their high percentages in the pockets and the RNA-binding pockets indicate their importance in the RNA-binding events.

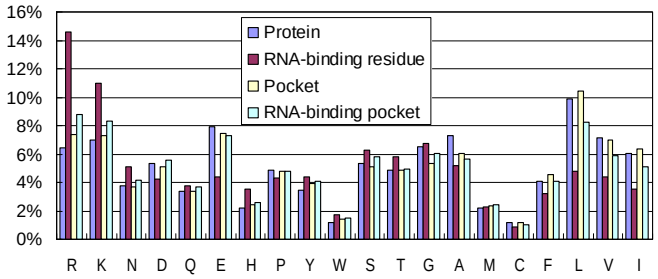

Fig. 1. The composition ratio of 20 amino acid residues in the 158 proteins, RNA-binding residues, pockets and the RNA-binding pockets which involved at least one RNA-binding residue. Several amino acid residues contain higher composition ratios in the RNA-binding events. The residues are positioned in the order of increasing hydrophobicity.

B. Classes of protein global structures and domains

The classes of protein structures and underlying domains provide the overview of structure similarities and functional relationships in RNA-binding. By employing CE algorithm [18] for global structure alignments in these RNA-binding proteins, we identify the protein structure similarities in an all-against-all manner. We use one node to represent one protein in the structure comparison. When the similarity Z-score between two proteins exceeds a given threshold of 3.7, they are linked by an edge. In such way, we build a protein similarity network as shown in Figure 2. We implement a fast community structure detection method in the protein similarity network to clustering these RNA-binding proteins into 8 groups. 15 isolated proteins which can not find their similar structure neighbors are also shown by gray nodes in Figure 2. The protein with highest degree in each group is identified as the structure representatives of these groups individually. As shown in Figure 2, protein '1B23:P' is the representative structure of the protein group colored in red and '2BH2:A' is the representative structure of the protein group colored in yellow. They are also defined as the hub proteins in each group. The global protein structures can be simply narrowed down to these representative structures by the network-based clustering method.

Note that the sequence similarities in these proteins are less than 25%, i.e., these proteins can not find their homologous relationships each other only by sequence alignments. While as shown in Figure 2, many significant similarity relationships are detected in these proteins and they are clustered as 8 major groups. This provides evidence that RNA-binding functions in proteins are more conserved in three-dimensional structure space than that in the sequence space. This inspires us to subsequently explore the domains underlying these proteins. Table I lists the Pfam domain annotations in these proteins, which generate protein families using hidden Markov models [19]. These proteins can be mainly classified in 27 Pfam superfamilies. Several domains still can not find their corresponding clans currently which are also listed in Table I. The diversity and complexity of the domains in the proteins indicate the functional complexity of protein-RNA binding events, such as 'RNA polymerase', 'tRNA synthetases', 'ssRNA viruses coating', 'tRNA-binding arm domain', 'Ribosomal RNA adenine dimethylase', 'RNA recognition motif', 'Ribonuclease', 'Zinc finger' and 'PAZ domain'. Most of the RNA-binding proteins (102/158) contain at least one annotated RNA-binding domains shown in Table I. Although the proteins can not find significant sequence similarities each other, some of them contain the same RNA-binding domains. For instance, protein '1EIY:B' and protein '2DU3:A' can not find the significant sequence similarities, but both of them contain the 'tRNA synthetases class II core' domain. The same domain underlying the two proteins determines their structure similarity shown in Figure 2 and they are also clustered into the same group colored in blue. This implies that the domain units are very important for performing RNA-binding functions in proteins. Several proteins such as '1EIY:B' contain several domains simultaneously. This indicates they will perform multiple roles in binding RNA. Different parts of protein structure perform different RNA-binding related functions. It is known that the RNA-binding domains specify the sequence profiles and patterns of local structure basis of protein-RNA recognition [1], [4]. And the results illustrate we should further analyze the local structures of RNA-binding pockets on protein surfaces to decipher their functional importance and complexity, and identify the RNA-binding structure motifs [7].

C. RNA-binding local structure motifs in proteins

We identify the pockets located on the protein surfaces by CASTp [16]. These pockets are extracted down from these RNA-biding proteins, among which the pairwise structure alignments are performed by SAMO [20]. The local structure alignment measurements are transformed into normalized Q-scores [21]. When the similarity score between two pockets exceeds a given threshold of 0.8, we link an edge between them. The built pocket similarity network describes the local structural similarities and relationships in these RNA-binding pockets. Based on the former identified protein groups, we relocate these pockets into the groups which contain their associate proteins. The network topology is shown in Figure 3. We also illustrate the structure of these pockets and their positions in the global structures of the representative proteins. Here, isolated pockets which can not find their similar structure neighbors are not shown. From Figure 3, we find the structure relationships in the RNA-binding pockets are not perfectly consistent with the global structure relationships in their associated proteins. For instance, protein '1A34:A' can not detect a significant global structure similarity with protein '1B23:P'. While in their contained RNA-binding pockets, pocket '1A34:A:9' is identified to be significantly similar to pocket '2BH2:A:89'. In the local binding structure space, the structure similarities related to RNA-binding functions become more concrete and detailed. In some isolated proteins of Figure 2, such as protein '163D:B' colored in gray, pockets '163D:B:7' and '163D:B:10' are identified to be similar with pocket '1B23:P:57', respectively. This demonstrates that the structure similarities underlying the local RNA-binding pockets are more complicated than that at the protein and domain levels.

In the pocket similarity network, we identify the hub

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

TABLE I. THE PFAM SUPERFAMILY AND DOMAINS IN THE RNA-BINDING PROTEINS. THE BOLDED PROTEINS ARE THOSE HUB PROTEINS IN THE IDENTIFIED STRUCTURE CLUSTERS.

Clan ID	Description	Domain ID	Description	Protein
CL0007	K-Homology (KH) domain Superfamily	PF00013	KH domain	1EC6:A; 2ANN:A; 3AEV:B
		PF07650	KH domain	3IEV:A
		PF15287	NusA-like KH domain	2ATW:A
CL0027	RNA dependent RNA polymerase	PF22924	Reverse transcriptase (RNA-dependent DNA polymerase)	1HYS:B
		PF30561	RNA dependent RNA polymerase	1UVL:A; 2E9T:A; 3BSO:A
CL0039	HUP - HIGH-signature proteins, UspA, and PP-ATPase	PF38198	tRNA synthetases class I (I, L, M and V)	1FFY:A; 1GAX:A; 2BTE:A; 2CSX:A
		PF00579	tRNA synthetases class I (W and Y)	1H3E:A; 1J1U:A; 2AKE:A
		PF00749	tRNA synthetases class I (E and Q), catalytic domain	1N78:A; 1QTQ:A
		PF00750	tRNA synthetases class I (R)	1F7U:A; 2ZUE:A
		PF01406	tRNA synthetases class I (C) catalytic domain	1U0B:B
		PF03054	tRNA methyl transferase	2DER:A
		PF09334	tRNA synthetases class I (M)	2BTE:A; 2CSX:A
		PF13603	Leucyl-tRNA synthetase, Domain 2	2BTE:A
CL0040	Class II aminoacyl-tRNA and Biotin synthetases	PF00152	tRNA synthetases class II (D, K and N)	1ASY:A; 1C0A:A
		PF00587	tRNA synthetase class II core domain (G, H, P, S and T)	1KOG:A; 1QF6:A; 1SER:A
		PF01409	tRNA synthetases class II core domain (F)	**1EIY:B**; 2DU3:A; 2DU4:A; 2IY5:A
CL0055	Positive stranded ssRNA viruses coat protein	PF00983	Tymovirus coat protein	1DDL:A
		PF01829	Peptidase A6 family	1F8V:A; 2Q23:C
		PF02247	Large coat protein	**1BMV:2**
CL0063	FAD/NAD(P)-binding Rossmann fold Superfamily	PF00398	Ribosomal RNA adenine dimethylase	3FTF:A
		PF01358	Poly A polymerase regulatory subunit	1AV6:A
		PF02475	Met-10+ like-protein	2ZZM:A
		PF05958	tRNA (Uracil-5-)-methyltransferase	**2BH2:A**; 3BT7:A
		PF13489	Methyltransferase domain	3HTX:A
CL0101	Pelota - RNA ribose binding superfamily	PF01248	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family	1E7K:A; 1SDS:A; 1T0K:B
CL0178	PUA/ASCH superfamily	PF01472	PUA domain	1J2B:A; 1R3E:A; 1ZE2:B; 2RFK:A
		PF09157	Pseudouridine synthase II TruB, C-terminal	1K8W:A
CL0196	DSRM-like clan	PF00035	Double-stranded RNA binding motif	1DI2:A; 1RC7:A; 3ADI:A; 3ADL:A
CL0219	Ribonuclease H-like superfamily	PF00075	RNase H	1ZBI:A; 1ZBL:A; 2QK9:A
		PF00929	Exonuclease	1ZBH:A
		PF02171	Piwi domain	**1YTU:A**; 2F8S:A; 3F73:A
CL0221	RRM-like clan	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	1B7F:A; 1CVJ:A; **1ZH5:A**; 2G4B:A; 3NNH:A
		PF13893		1A9N:B
		PF14259		2G4B:A
CL0250	GAD domain superfamily	PF02938	GAD domain	1C0A:A; 2D6F:C
CL0258	DALR superfamily	PF05746	DALR anticodon binding domain	1F7U:A; 2ZUE:A
		PF09190	DALR domain	1U0B:B
CL0260	Nucleotidyltransferase superfamily	PF01743	Poly A polymerase head domain	1VFG:A
		PF01909	Nucleotidyltransferase domain	2DR8:A; 2Q66:A
CL0298	tRNA-binding arm superfamily	PF02912	Aminoacyl tRNA synthetase class II, N-terminal domain	2IY5:A
		PF10458	Valyl tRNA synthetase tRNA binding arm	1GAX:A
CL0303	Helix-two-turns-helix superfamily	PF00416	Ribosomal protein S13/S18	1XMQ:M
CL0329	Ribosomal protein S5 domain 2-like superfamily	PF01138	3' exoribonuclease family, domain 1	2JEA:A; 3M7N:F
CL0383	Phenylalanine- and lysidine-tRNA synthetase domain superfamily	PF03483	B3/4 domain	**1EIY:B**
CL0410	LEF-8 like region of RNA polymerase Rpb2	PF00562	RNA polymerase Rpb2, domain 6	2NVQ:B
CL0441	RNA-DNA binding Alba-like superfamily	PF12328	Rpp20 subunit of nuclear RNase MRP and P	3IAB:B
CL0458	Class II aaRS Anticodon-binding domain-like	PF03129	Anticodon binding domain	1KOG:A; 1QF6:A
CL0476	tRNA-intron endonuclease catalytic domain-like N-term	PF02778	tRNA intron endonuclease, N-terminal domain	2GJW:A
CL0480	Ribosomal L1 protein superfamily	PF00687	Ribosomal protein L1p/L10e family	1MZP:A; 2VPL:A
CL0492	S4 domain superfamily	PF01479	S4 domain	3DH3:A
CL0527	Sm (Small RNA binding protein domain)	PF01423	LSM domain	**1KQ2:A**; 1M8V:A; 3AHU:A
CL0537	CCCH-zinc finger	PF00642	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	3D2S:A
CL0539	RNase III domain-like superfamily	PF14622	Ribonuclease-III-like	1RC7:A
–	no clan	PF03143	Elongation factor Tu C-terminal domain	**1B23:P**
–	no clan	PF09021	HutP (Histidine utilizing Protein)	**1WPU:A**
–	no clan	PF09107	Elongation factor SelB, winged helix	1WSU:A; 2PJP:A; **2PLY:A**
–	no clan	PF02170	PAZ domain	1SI3:A

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

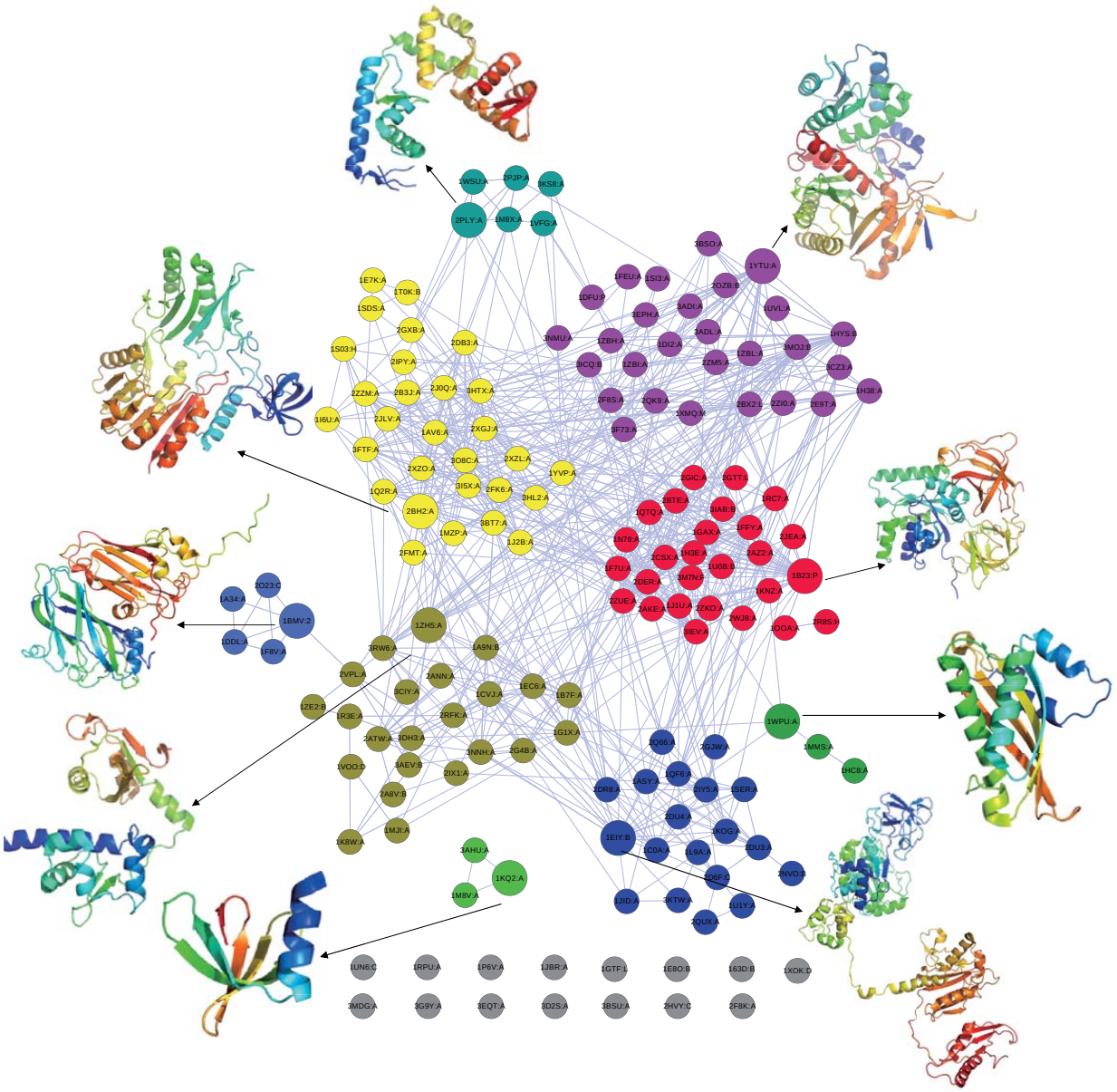

Fig. 2. The protein similarity network in the RNA-binding proteins. The representative protein structures are shown in bigger node size and their structures are also displayed simultaneously. The gray nodes are those isolated proteins which can not find their structure neighbors with significant similarity. The networks in this paper are illustrated by Cytoscape (www.cytoscape.org) and the structures are illustrated by PyMol (www.pymol.org).

pockets with highest degrees as the RNA-binding structure motifs. They are local structure centers and templates in these RNA-binding pockets. As shown Figure 3, these local structure motifs are shown with their location in their associated proteins. Combined with the domains listed in Table I, we find most of these RNA-binding structure motifs locate in the RNA-binding domains. This demonstrates that the RNA-binding domains fold to certain local structure patterns from the three-dimensional perspective and then the formed pockets on protein surfaces facilitate the local spots and environments in the request of binding RNA. The pockets are the local structure motifs of RNA-binding. The major groups of these RNA-binding pockets are also identified by the similarity network framework. The protein-RNA local structure binding motifs are then identified and extracted.

For illustration purpose, Figure 4 shows the information of two proteins and their two pockets, respectively. Protein '1ASY:A' and '1B23:P' can not detect their significant similarity from their sequences (sequence identity is 16.3%) and structures (CE alignment Z-score is 3.1). While pocket '1ASY:A:75' and pocket '1B23:P:35' can find the local structure similarities each other. The positions of the two pockets in the two sequences and structures are also shown in Figure 4, respectively. Pocket '1ASY:A:75' locates in the domain 'PF00152: tRNAsynt_2' in protein '1ASY:A' and pocket '1B23:P:35' locates in the domains of 'PF00009:GTP_EFTU' and 'PF03144:GTP_EFTU_D2' in protein '1B23:P'. Although the two proteins can not find their global sequence and structure similarities, the contained similar pockets provide the detailed local structures needed for binding RNA. Moreover,

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

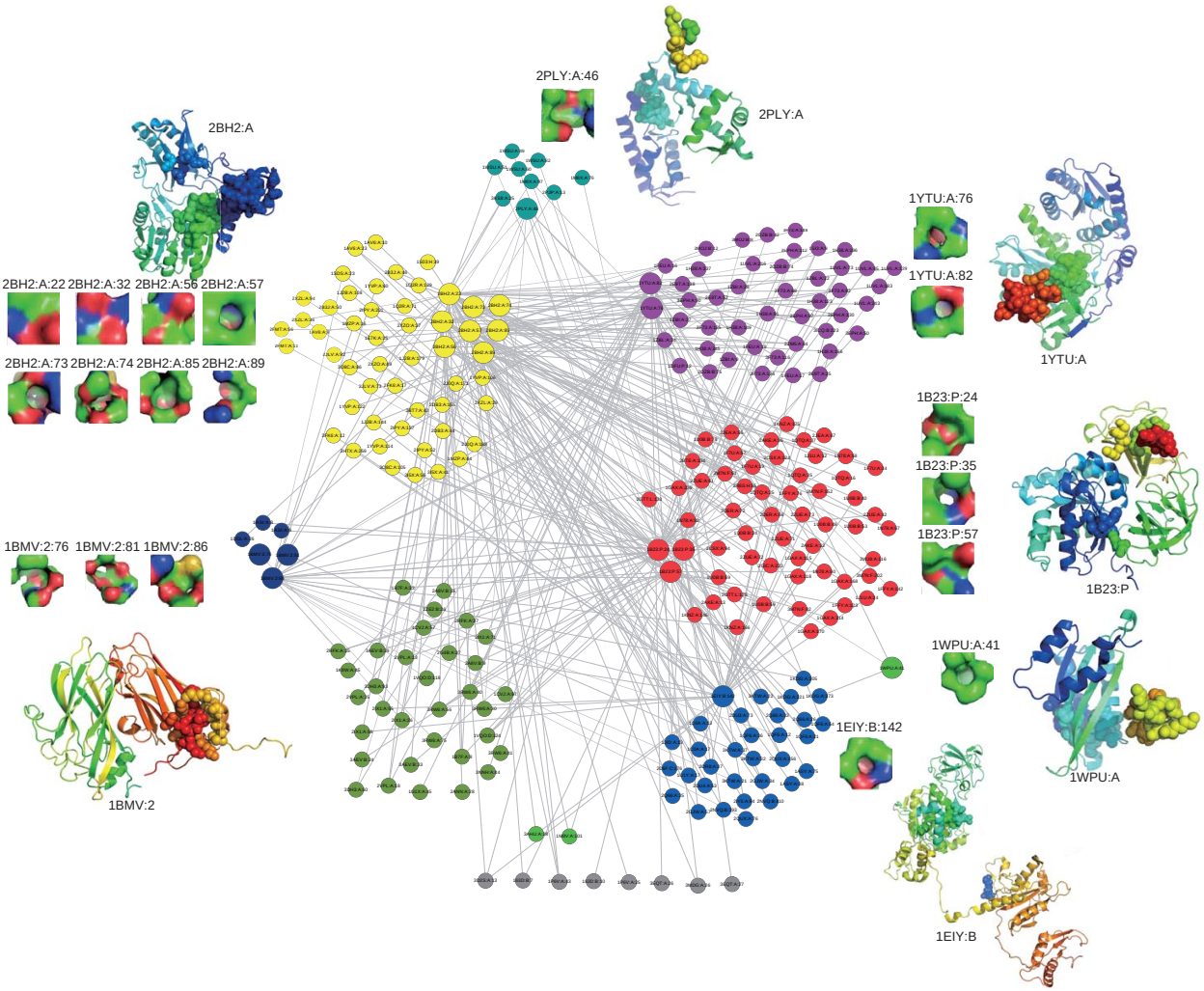

Fig. 3. The pocket similarity network in the RNA-binding pockets, which are affiliated in the RNA-binding proteins shown in the former Figure 2. The identified local structure motifs of RNA-binding pocket and the locations of these motifs in their involved proteins are also shown.

the domains are also different each other in the two proteins. It implies that these local structures on protein surfaces determine the concrete RNA-binding functions in the two proteins respectively. The results also indicate clearly that the identification of RNA-binding motifs should extract the structure patterns in these local pockets, instead of from global protein structures and domains. The pockets shown in Figure 3 and Figure 4 are these identified RNA-binding structure motifs in our collected RNA-binding proteins.

III. Conclusion

In this paper, we proposed a systematic identification of local structure binding motifs of protein-RNA recognition to reveal the local structural basis and patterns in the RNA-binding events. Firstly, we found that most of the protein-RNA binding events take place at the pockets on protein surfaces. The major binding pockets classes and the classes of their associated protein structures and domains are also identified. The findings provide evidences for the importance of local binding pockets in the protein-RNA recognitions. The classified binding pockets and their structure motif patterns will benefit the RNA-interaction-related proteins design and engineering.

IV. Materials and Methods

A. Datasets

We compile the available protein-RNA interactions from PDB [22]. We download the documented protein-RNA complexes from the PDB database and extract 896 protein-RNA complexes. We eliminate the complexes when their protein sequence lengths are not within from 15 to 200 amino acid residues and RNA sequence lengths are not within from 5 to 200 nucleotide residues. After removing the homologous proteins by the sequence similarity of 25% and RNAs by the sequence similarity of 60% via BLASTclust [23], we achieve 158 non-redundant protein-RNA complexes.

The interacting residues of protein and RNA are identified by ENTANGLE [24]. Hydrogen bonding, electrostatic, hydrophobic and van der Waals interactions between protein and RNA are considered as the types of protein-RNA interactions. Pockets are empty concavities on a protein surface into which

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

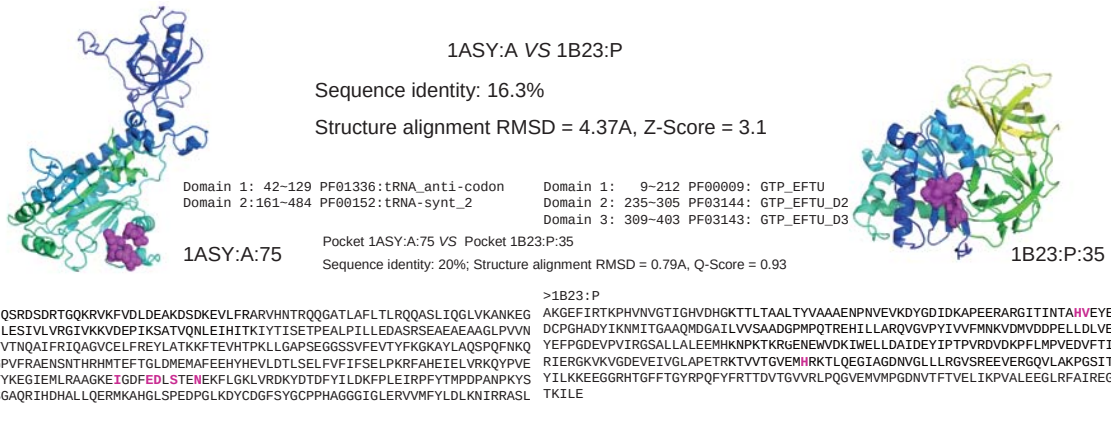

Fig. 4. The information of proteins '1ASY:A' and '1B23:P'. The locations of pockets '1ASY:A:75' and '1B23:P:35' in the protein sequences and structures are also shown respectively.

solvent (probe sphere 1.4) can gain access, i.e., these concavities have mouth openings connecting their interior with the outside bulk solution. The surface accessible pockets of proteins are identified by CASTp [16], which is based on recent theoretical and algorithmic results of computational geometry. The pockets are identified from the protein surfaces of these complexes respectively, and their position and coordination are extracted accordingly. The RNA-binding pockets are defined as these pockets containing at least one RNA-binding residue individually.

B. Clustering of global and local structures via similarity networks

The similarity between proteins and that between pockets are measured by structure alignment algorithms CE [18] and SAMO [20], respectively. CE outputs a statistical significance of Z-score to measure the global structure similarity between two proteins. SAMO determines the alignment metrics of $RMSD$ (root mean square deviation) and the aligned residue number N_{align}. The alignment quality measurements are then transformed into a Q-score, i.e., $Q = N_{align}^2/[1 + (RMSD/R_0)^2]N_1N_2$, where R_0 is a normalizing factor (set as 3.0), and N_1 and N_2 refer to the sequence lengths of the two pockets [21]. We build a protein similarity network and a pocket similarity network for describing the global and local structure similarities in these proteins and in their RNA-binding pockets, respectively. The nodes in the networks are these proteins and pockets. When the Z-score between two proteins exceeds 3.7, we link an edge between them for constructing the protein similarity network. Similarly, when the Q-score between two pockets exceeds 0.8, we add an edge between them for the pocket similarity network. We implement the clustering processes based on the similarity network model [25]. By employing a network community detection algorithm [26], we decompose the protein similarity network into several protein groups. The protein groups and then the pocket clusters identified from the similarity network framework are the similar global and local structures related to RNA binding respectively.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 31100949 and the Fundamental Research Funds of Shandong University under Grant No. 2014TB006.

REFERENCES

[1] B. M. Lunde, C. Moore, and G. Varani, "RNA-binding proteins: modular design for efficient function," Nat. Rev. Mol. Cell Biol., vol. 8, pp. 479-490, 2007.

[2] S. Jones, D. T. Daley, N. M. Luscombe, H. M. Berman, and J. M. Thornton, "Protein-RNA interactions: a structural analysis," Nucleic Acids Res., vol. 29, pp. 943-954, 2001.

[3] D. J. Hogan, D. P. Riordan, A. P. Gerber, D. Herschlag, and P. O. Brown, "Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system," PLoS Biol., vol. 6, p. e255, 2008.

[4] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," FEBS Lett., vol. 582, pp. 1977-1986, 2008.

[5] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," Nat. Rev. Genet., vol. 5, pp. 522-531, 2004.

[6] A. Shulman-Peleg, M. Shatsky, R. Nussinov, and H. J. Wolfson, "Prediction of interacting single-stranded RNA bases by protein-binding patterns," J. Mol. Biol., vol. 379, pp. 299-316, 2008.

[7] Z. P. Liu, L. Y. Wu, Y. Wang, L. Chen, and X. S. Zhang, "Predicting gene ontology functions from protein's regional surface structures," BMC Bioinformatics, vol. 8, p. 475, 2007.

[8] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Bridging protein local structures and protein functions," Amino Acids, vol. 35, pp. 627-650, 2008.

[9] J. J. Ellis, M. Broom, and S. Jones, "Protein-RNA interactions: structural analysis and functional classes," Proteins, vol. 66, pp. 903-911, 2007.

[10] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," Bioinformatics, vol. 26, pp. 1616-1622, 2010.

[11] M. Terribilini, J. D. Sander, J. H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs, "RNABindR: a server for analyzing and predicting RNA-binding sites in proteins," Nucleic Acids Res., vol. 35, pp. W578-584, 2007.

[12] L. Wang and S. J. Brown, "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," Nucleic Acids Res., vol. 34, pp. W243-248, 2006.

[13] Y. Wang, Z. Xue, G. Shen, and J. Xu, "PRINTR: prediction of RNA binding sites in proteins using SVM and profiles," Amino Acids, vol. 35, pp. 295-302, 2008.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

[14] Y. Wang, X. Chen, Z. P. Liu, Q. Huang, D. Xu, X. S. Zhang, R. Chen, and L. Chen, "De novo prediction of RNA-protein interactions from sequence information," Mol. Biosyst., vol. 9, pp. 133-142, 2013.

[15] U. K. Muppirala, V. G. Honavar, and D. Dobbs, "Predicting RNA-protein interactions using only sequence information," BMC Bioinformatics, vol. 12, p. 489, 2011.

[16] T. A. Binkowski, S. Naghibzadeh, and J. Liang, "CASTp: Computed Atlas of Surface Topography of proteins," Nucleic Acids Res., vol. 31, pp. 3352-3355, 2003.

[17] M. Terribilini, J. H. Lee, C. Yan, R. L. Jernigan, V. Honavar, and D. Dobbs, "Prediction of RNA binding sites in proteins from amino acid sequence," RNA, vol. 12, pp. 1450-1462, 2006.

[18] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," Protein Eng., vol. 11, pp. 739-747, 1998.

[19] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: the protein families database," Nucleic Acids Res., vol. 42, pp. D222-230, 2014.

[20] L. Chen, L. Y. Wu, Y. Wang, S. Zhang, and X. S. Zhang, "Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison," BMC Struct. Biol., vol. 6, p. 18, 2006.

[21] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," Acta Crystallogr. D Biol. Crystallogr., vol. 60, pp. 2256-2268, 2004.

[22] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Res., vol. 28, pp. 235-242, 2000.

[23] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res., vol. 25, pp. 3389-3402, 1997.

[24] J. Allers and Y. Shamoo, "Structure-based analysis of protein-RNA interactions using the program ENTANGLE," J. Mol. Biol., vol. 311, pp. 75-86, 2001.

[25] Z. P. Liu, L. Y. Wu, Y. Wang, and X. S. Zhang, "Protein cavity clustering based on community structure of pocket similarity network," Int. J. Bioinform. Res. Appl., vol. 4, pp. 445-460, 2008.

[26] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, p. 066111, 2004.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE