SCIENTIFIC REPORTS

OPEN

An account of solvent accessibility in protein-RNA recognition

Sunandan Mukherjee & Ranjit Prasad Bahadur

Protein–RNA recognition often induces conformational changes in binding partners. Consequently, the solvent accessible surface area (SASA) buried in contact estimated from the co-crystal structures may differ from that calculated using their unbound forms. To evaluate the change in accessibility upon binding, we compare SASA of 126 protein-RNA complexes between bound and unbound forms. We observe, in majority of cases the interface of both the binding partners gain accessibility upon binding, which is often associated with either large domain movements or secondary structural transitions in RNA-binding proteins (RBPs), and binding-induced conformational changes in RNAs. At the non-interface region, majority of RNAs lose accessibility upon binding, however, no such preference is observed for RBPs. Side chains of RBPs have major contribution in change in accessibility. In case of flexible binding, we find a moderate correlation between the binding free energy and change in accessibility at the interface. Finally, we introduce a parameter, the ratio of gain to loss of accessibility upon binding, which can be used to identify the native solution among the flexible docking models. Our findings provide fundamental insights into the relationship between flexibility and solvent accessibility, and advance our understanding on binding induced folding in protein-RNA recognition.

Protein-RNA recognition is essential for gene expression and its regulations. The initial contact between a RNA-binding protein (RBP) and a RNA, often termed as encounter complex, triggers subsequent conformational changes in order to form a stable and functional association[1,2]. These conformational changes can either be of small scale including side chain movements of amino acid residues or base flipping of nucleotides, or be of large scale movements such as reorientation of polypeptide domains or change in secondary structures of RNA. Moreover, secondary structural transitions can also induce major conformational changes in both the binding partners[3]. It has been observed that the conformational changes upon binding are often associated with significant changes in solvent accessibility in the binding partners. Lee and Richard, in 1971[4], first coined the term "accessible surface area" to quantify the area of protein surface. Later, Chothia[5,6] described the correlation between accessible surface area and hydrophobic energy in protein folding. According to his study, the gain in ΔG per squared Angstrom decrease in solvent accessible surface area (SASA) of proteins is 25 cal/mol. Recent studies have shown that relative solvent accessibility can be used to predict the extent of conformational changes associated with protein-protein recognition[7]. Besides, it has also been found that the bound conformations of macromolecules have larger SASA than their unbound states[8]. Moreover, the intrinsic flexibility of proteins can also be measured by their buried and accessible surface area[9]. Recently, Barik *et al*.[10] showed that the change in SASA upon binding can be used as a parameter to predict the binding hotspots at protein–RNA interfaces. The prerequisite to study the change in solvent accessibility upon protein-RNA binding is the atomic structures of the complexes and their corresponding unbound forms of the binding partners. The growing interests to decipher the 3-dimensional structures of protein-RNA complexes and their unbound structures facilitated the development of protein-RNA docking benchmarks[11–13].

In this study, we evaluate the change in SASA values calculated from the bound complex and their corresponding unbound components of protein-RNA complexes taken from the docking benchmark version 2[13]. We find, in majority of the cases the interface of both the binding partners gain accessibility in order to provide more surface area to promote the stable interactions. However, majority of RNA non-interface region lose accessibility, while, no such significant bias is observed at the non-interface region of RBPs. The change in interface accessibility is significantly contributed by the side chains, however, a moderate correlation between the change in accessibility and the backbone conformation is also observed. Interestingly, large change in accessibility is observed when the binding is more flexible including large domain movements and secondary structural transition of

Computational Structural Biology Laboratory, Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur, 721302, India. Correspondence and requests for materials should be addressed to R.P.B. (email: r.bahadur@hijli.iitkgp.ernet.in)

	tRNA (A)	rRNA (B)	duplex RNA (C)	single-stranded RNA (D)	All
No of complexes	28	5	40	53	126
$P_U R_U$	10	3	4	4	21
$P_U R_B$	16	2	32	45	95
$P_B R_U$	2	0	4	4	10
Average B (Å²)					
B^B	2625	1733	2329	1891	2187
Average ASA (Å²)					
B_P^B	1259 (±566)	839 (±276)	1153 (±681)	863 (±580)	1040 (±626)
A_P^B	23783 (±11457)	8282 (±2342)	18598 (±11530)	16912 (±11836)	18559 (±11903)
B_R^B	1375 (±736)	965 (±198)	1007 (±284)	1048 (±604)	1132 (±572)
A_R^B	8028 (±3363)	6553 (±295)	7701 (±6721)	4786 (±4042)	6943 (±4922)
Average δA_P (Å²)[a]					
Interface	−132.4 (±174.3)	37.5 (±62.8)	−145.9 (±213.7)	−111.5 (±192.1)	−120.5 (±195.2)
Exposed	−174.3 (±167.9)	−23.1 (±20.9)	−167.3 (±209.4)	−183.4 (±169.5)	−172.0 (±184.3)
Buried	43.5 (±35.5)	77.8 (±46.9)	88.8 (±74.0)	87.4 (±75.6)	77.2 (±68.0)
Non-interface	4.5 (±35.3)	33.3 (±20.7)	−10.3 (±31.5)	9.8 (±46.6)	3.4 (±40.1)
Exposed	−25.2 (±16.3)	−5.5 (±NA)	−31.0 (±19.1)	−19.3 (±27.5)	−24.6 (±24.1)
Buried	23.0 (±27.7)	59.2 (±16.7)	26.5 (±24.1)	33.5 (±49.0)	30.4 (±33.5)
Average δA_R (Å²)[b]					
Interface	−141.0 (±104.3)	−73.9 (±195.2)	16.1 (±135.9)	−174.6 (±67.0)	−92.5 (±143.7)
Exposed	−158.9 (±94.3)	−184.4 (±143.3)	−68.0 (±32.1)	−174.6 (±67.0)	−144.1 (±91.8)
Buried	20.0 (NA)	146.9 (NA)	142.3 (±134.6)	0.0 (NA)	122.7 (±119.1)
Non-interface	16.5 (±40.1)	32.6 (±18.3)	77.9 (±112.2)	26.1 (±41.5)	40.3 (±75.8)
Exposed	−29.7 (±34.0)	0.0 (±NA)	−20.3 (±12.2)	−30.2 (±6.1)	−25.6 (±21.9)
Buried	36.3 (±22.3)	32.6 (±18.3)	143.3 (±100.8)	44.8 (±29.9)	67.3 (±73.5)

Table 1. Statistics on change in accessibility upon binding in protein–RNA complexes. [a]Values are calculated on 116 $P_U R_U$ and $P_U R_B$ complexes. [b]Values are calculated on 31 $P_U R_U$ and $P_B R_U$ complexes. Standard deviations are provided in parentheses.

RBPs upon binding. We find a moderate correlation between the change in accessibility and binding free energy when the interface undergoes significant change in conformation upon binding. Analysis of secondary structural elements reveals that loop-to-helix and helix-to-loop transitions upon binding gain significant accessibility at the interface. Additionally, we show the amino acid residues and nucleotides that do not participate in intermolecular hydrogen bonds (H-bonds) undergo significant change in accessibility upon binding compared to those which are involved in such interactions. We have also investigated the preference of amino acid residues and nucleotides to lose accessibility (buried) or gain accessibility (exposed) upon binding. Finally, we introduce a new parameter, the ratio of gain to loss of solvent accessibility upon binding (GL ratio), which can be effectively used to score the flexible docking solutions to identify the near native structure. These findings have significant applications in designing flexible protein-RNA docking algorithms and engineering protein-RNA interfaces.

Results

Dataset of bound and unbound structures.
The dataset consists of 126 protein-RNA complexes for which at least one interacting partner is available in the unbound form. Of these 126 complexes, 28 are in class A, 5 are in class B, 40 are in class C and 53 are in class D (refer to Materials and Methods section and Supplementary Table S1). Based on their availability in the unbound form, we find 21 are of $P_U R_U$ type, where both the protein and the RNA are available in the unbound form, 95 are of $P_U R_B$ type, where only the protein is available in the unbound form, and 10 are of $P_B R_U$ type, where only the RNA is available in the unbound form (Table 1). Local alignment of the polypeptide chains between the unbound and the bound (U/B) structures reveals that 93 out of 116 have sequence identity >98%, while the rest have values between 90% and 98%. On the other hand, sequence identity of 20 out of 31 U/B pairs of polynucleotides have values >98% and the rest have values between 90% and 98%. We have discarded 896 (~2%) residues and 96 (~6.7%) nucleotides in the entire dataset due to the mismatch in the alignment between U/B pairs.

Change in accessibility at the protein-RNA interfaces.
The overall change in accessibility upon binding is a cumulative effect of many local conformational rearrangements. Some residues get exposed by burying the others or vice-versa. Change in accessibility of the interface atoms upon binding was calculated by comparing their SASA values in bound and unbound states. On an average, RBPs gain 120.5 Å² of solvent accessibility at the interface upon binding with RNA (Table 1). We find in 92 out of 116 cases, interface region of RBPs gain accessibility upon binding with an average δA_P^{int} (refer to Materials and Methods section) of −172.0 Å². In the remaining 24 cases, positive changes in δA_P^{int} are observed with an average of 77.2 Å², indicating a loss in accessibility at the
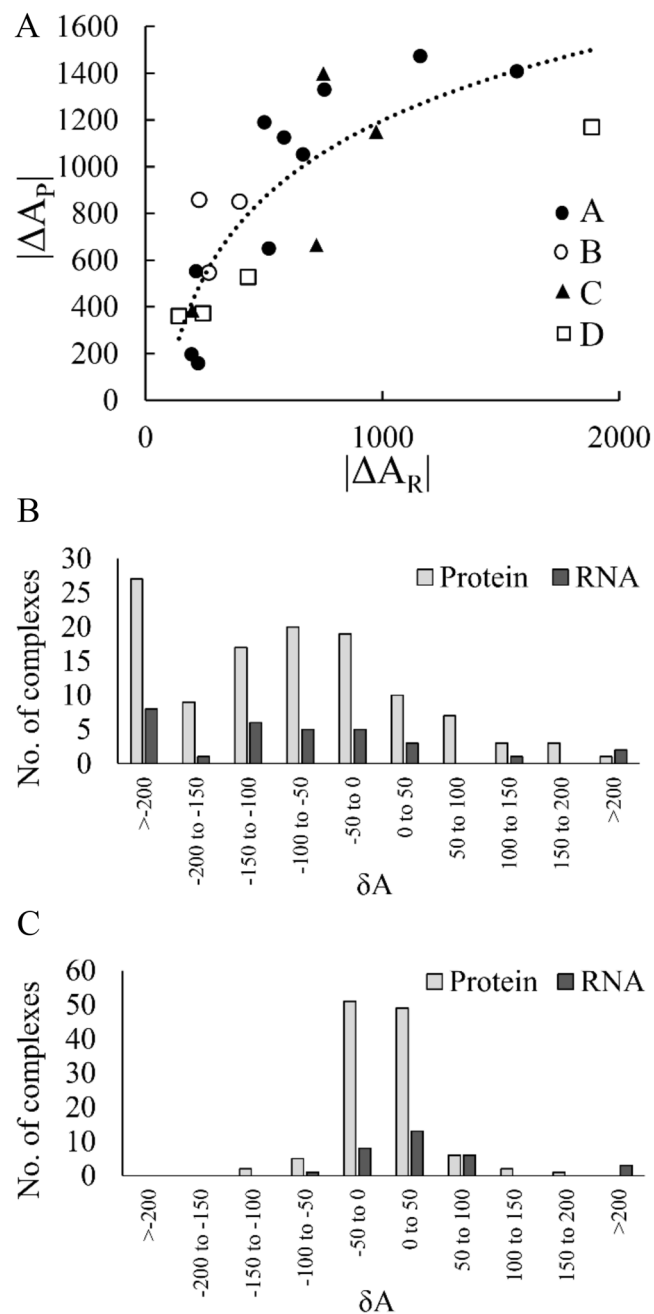
Figure 1. Distribution of change in accessibility in RBPs and RNAs upon binding. **(A)** Correlation between $|\Delta A_P|$ and $|\Delta A_R|$ at the protein-RNA interfaces for 21 UU cases. The different classes of complexes are shown in different symbols. Distributions of δA in 116 RBPs and in 31 RNAs at the protein-RNA **(B)** interface and **(C)** non-interface regions.

interface. On an average, interface region of RNAs gain 92.5 Å^2 of solvent accessibility upon binding with RBPs. Majority of them, 80% (25 out of 31), show negative δA_R^{int} with an average of -144.1 Å^2 (Table 1). Remaining, only 20%, show positive changes with an average of 122.7 Å^2, indicating a loss of accessibility.

We used 21 $P_U R_U$ cases to quantify the overall change in the accessibility at the interface when both RBPs and their partner RNAs are available in the unbound form. Here, we find an average δA_{P+R}^{int} (refer to Materials and Methods section) of -221.2 Å^2. In this subset, only four complexes lose accessibility upon binding (average $\delta A_{P+R}^{int} = 39.2$ Å^2, range is from 22.2 Å^2 to 59.5 Å^2), of which two are from class B and one each from class A and class D. In rest of the 17 cases, the change is negative with an average δA_{P+R}^{int} of -282.5 Å^2 (range is from -14.3 Å^2 to -1285 Å^2). A significant correlation is observed between ΔA_P^{int} and ΔA_R^{int} (R$^2 = 0.7$) in the entire dataset (Fig. 1A).
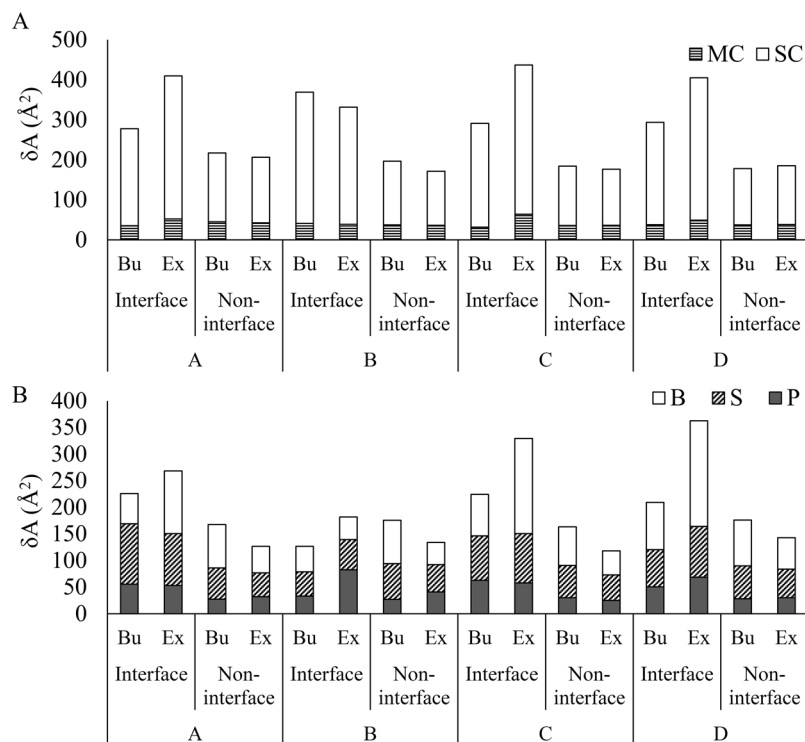
Figure 2. Distribution of δA in main chain and side chain calculated on 116 RBPs (**A**), and in phosphate, sugar and bases calculated on 31 RNAs (**B**). The average values are presented for Buried (Bu) and Exposed (Ex) surfaces in interface and in non-interface regions of different class of complexes.

Change in accessibility at the non-interface region. We have estimated the change in accessibility of amino acid residues and nucleotides at the non-interface region. Here, the average change in accessibility of RBPs is only 3.4 Å^2 (Table 1), which is significantly lower than that of the interface region. In the entire dataset, 50% RBPs show negative changes with an average of -24.6 Å^2, and 50% show positive changes with an average of 30.4 Å^2. At the non-interface region of the RNA, the average change in accessibility is 40.3 Å^2. In the entire dataset, majority (71%) of the RNAs lose accessibility upon binding with an average $\delta A_R^{non\text{-}int}$ of 67.3 Å^2. Only nine RNAs (29%) show negative changes with an average of -25.6 Å^2.

The distributions of δA$_P$ and δA$_R$ at the interface and at the non-interface regions are shown in Fig. 1B and 1C, respectively. At the non-interface region, the majority (86%) of $\delta A_P^{non\text{-}int}$ remain within the range between -50 Å^2 and 50 Å^2 (Fig. 1C). We find, irrespective of different classes, the change in accessibility at the interface region is always higher than the non-interface region (Fig. 2A). Buried (Bu: gain of accessibility) and exposed (Ex: loss of accessibility) surfaces of RBPs contribute almost equally to δA$_P$ in the non-interface region. However, at the interface, exposed surface contribute 60% to δA_P^{int}, whereas, the buried surface contributes only 40%. Therefore, the effective change remains higher at the interface compared to the non-interface region. About 80% of changes in δA$_P$ both at the interface and the non-interface regions are contributed by the side chain atoms (Fig. 2A). The change in δA_P^{int} at the interface is highest in class D followed by class B, class A and class C. Except in class B, significant difference is observed between the exposed and the buried surfaces in all other class of interfaces. The δA$_R$ values are always higher at the interface compared to the non-interface regions in the entire dataset as well as in different classes (Fig. 2B). At the interface, δA$_R$ is the highest for the bases followed by the sugar and the phosphate.

Effect of conformational change on accessibility. Conformational changes between unbound and bound forms are estimated in terms of *i*-rmsd, which is the root mean squared deviation of interface C$_\alpha$ and P atoms of amino acids and nucleotides, respectively. Based on the degree of conformational changes, the protein-RNA binding can be classified into rigid body (*i*-rmsd < 1.5 Å), semi-flexible (*i*-rmsd within 1.5 Å to 3.0 Å) and full flexible (*i*-rmsd > 3.0 Å)[11,13]. Although we find the average change in δA_P^{int} is -96 Å^2 and -100.4 Å^2 for rigid-body and semi-flexible bindings, respectively, the change is significantly higher (-248 Å^2) for full flexible binding. We find a moderate correlation (R = 0.6) between δA_P^{int} and *i*-rmsd. Besides, we also find the change in interface accessibility is significantly contributed by the side chain conformations (Fig. 2A), which is ignored in *i*-rmsd calculation. This can be exemplified in Fig. 3A,B, where the tRNA splicing endonuclease undergoes rigid body association (*i*-rmsd is 1.0 Å), however, its interface shows a significant change in accessibility (δA_P^{int} is -410.7 Å^2) upon binding with its partner RNA. Here, the side chain (δA_P^{int} is -356 Å^2) accounts for the large change in accessibility than its main chain (δA_P^{int} is -54.6 Å^2). Counter examples are also observed, where
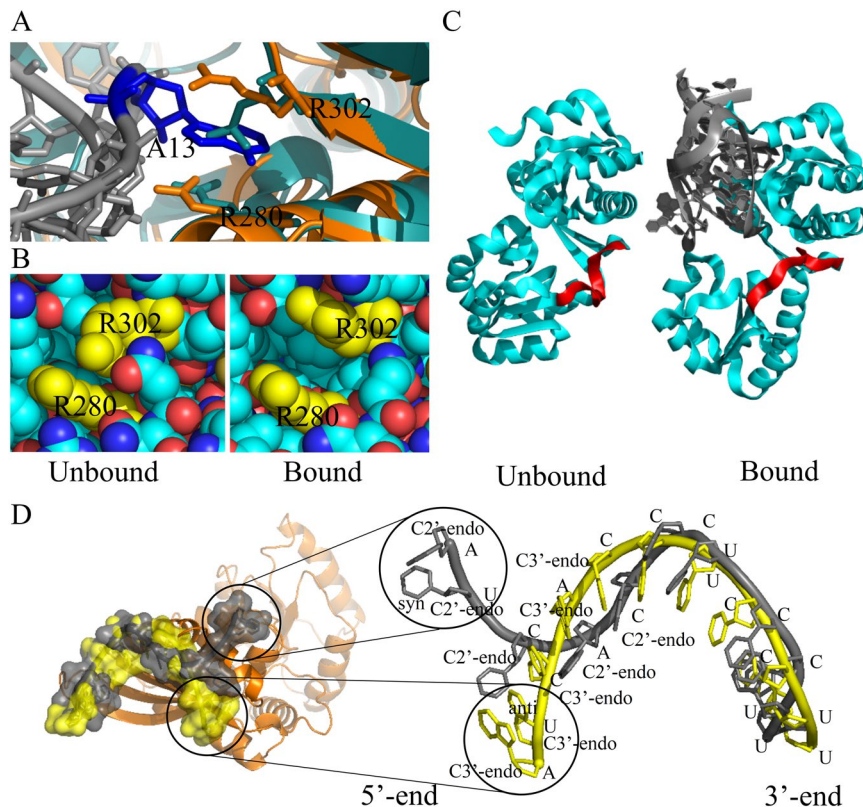
Figure 3. Change in accessibility on local and global conformational change. **(A)** Superposed structures of RNA splicing endonuclease in bound[42] (in orange, PDB id: 2GJW) and in unbound[43] (in cyan, PDB id: 1R0V) conformations with the RNA (shown in grey). Arg-nucleotide-Arg sandwich at the cleavage site of the nuclease is shown. Both the Arg are labeled and shown in stick. Change in conformation of R302 allows A13 (in blue) to protrude into the endonuclease pocket and stacked by the two Arg. **(B)** Top view of the aforementioned structure in sphere representation. Both the Arg are shown in yellow. In unbound structure, the endonuclease pocket is not accessible to the nucleotide. Change in conformation of R302 makes the pocket more accessible. **(C)** Unbound[44] (PDB id: 1AD2) and bound[45] (PDB id: 2HW8) structures of ribosomal protein L1 (in cyan). The loop at the hinge region connecting two domains is colored in red. RNA molecule in the bound structure is shown in grey cartoon. **(D)** Superposed structures of unbound (PDB id: 1SDR, in yellow) and bound (PDB id: 3IEV, in grey) forms of 12-nucleotides long 3′-end of 16 S rRNA with ERA. Protein is represented in orange cartoon.

the small change in interface accessibility does not correlate with the high i-rmsd values. This is exemplified in ribosomal L1 protein, which undergoes significant conformational change (i-rmsd is 5.1 Å) upon binding with its partner RNA even though the change in accessibility is only −2.2 Å^2. The N- and C-terminal domains of L1 are linked by a short and a long loop (Fig. 3C). In the unbound form, the buried surface area between these two domains is very small. Upon binding with RNA, the long loop acts as a hinge and moves both the domains apart to facilitates the RNA binding. This domain movement leads to higher i-rmsd without affecting the overall change in accessibility. Similarly, changes in accessibility may also be attributed to the backbone as well as to the conformational changes of sugar and bases of RNA. For an instance, *E. coli* Ras-like protein (ERA), which acts as a chaperone for folding and maturation of 16S rRNA induces a large conformational change in 12-nucleotides long 3′-end of 16 S rRNA. The RNA adopts a Z-like structure upon binding with the KH domain of ERA[14], and the estimated δA_R^{int} is −311.7 Å^2. The second U from the 5′-end of the 12-nucleotides sequence changes the conformation of the base (anti-to-syn) and the sugar pucker (C2′-endo-to-C3′-endo), and contributes −96.5 Å^2 change in accessibility (Fig. 3D).

Changes in secondary structural elements in RBPs upon binding. Conformation changes can alter the secondary structures during unbound to bound transition leading to the change in accessibility. Figure 4A shows the average $|\Delta A_P|$ accounts for different types of transitions in the secondary structural elements upon binding. We find the average change in accessibility at the interface is highest ($|\Delta A_P| = 47.5$ Å^2) in transitions from loop-to-helix followed by in transitions from helix-to-loop ($|\Delta A_P| = 41$ Å^2) and from loop-to-sheet ($|\Delta A_P| = 38.4$ Å^2). Figure 4B shows an example of loop-to-helix transition where the unstructured α1-helix of L25 protein in the unbound state (PDB id: 1B75) adopts the helical conformation upon binding with the major groove of 5 s rRNA (PDB id: 1DFU)[15]. The α1-helix loses 230 Å^2 upon binding with its partner RNA. We did not find any transition from helix-to-sheet or vice-versa at the interface.
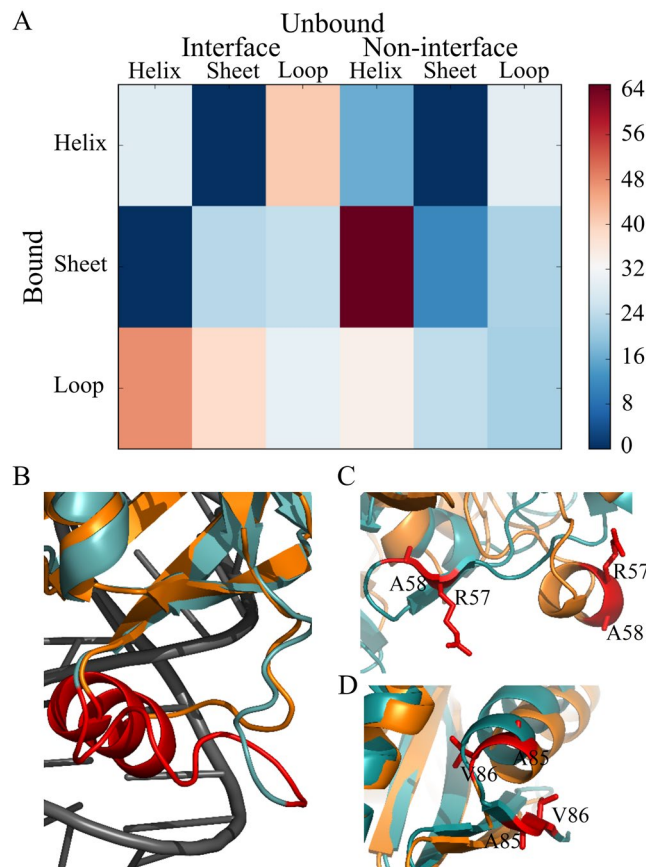
Figure 4. Changes in $|\Delta A|$ due to the transitions of secondary structural elements in RBPs upon binding with RNA. **(A)** Average $|\Delta A|$ calculated per transition is presented for both interface and non-interface regions. **(B)** A loop-to-helix transition. Here, the $\alpha1$-helix of L25 (Lys14 to Ala23, coloured in red) is unstructured in the unbound state (PDB id: 1B75), which adopts a helical conformation upon binding with the major groove of 5 s rRNA (PDB id: 1DFU). **(C)** A sheet-to-helix transition. Here, Arg57 and Ala58 (shown in red stick) of translation elongation factor EF-Tu are in sheet conformation in the unbound state (PDB id: 1TUI), which adopt helical conformations upon binding with the tRNA (Cys) (PDB id: 1B23). **(D)** Another example of sheet-to-helix transition. Here, Ala85 and Val86 (shown in red stick) of CCA-adding enzyme are in β-sheet conformations in the unbound state (PDB id: 1UET) of the enzyme, which adopt α-helical conformations upon binding with the t-RNA (PDB id: 2DRB). In all these figures, the protein in bound and unbound states is shown in orange and teal, respectively, and the RNA is shown in grey.

At the non-interface region, the highest change in accessibility is observed in transitions from sheet-to-helix ($|\Delta A_P| = 64.9$ Å^2). This change is observed in the following four residues from two different RBPs. Two residues, Arg57 and Ala58 in translation elongation factor EF-Tu (PDB id: 1TUI), undergo sheet-to-helix transitions upon binding with the tRNA(Cys) (PDB id: 1B23) (Fig. 4C). The other two residues, Ala85 and Val86 in the unbound state of the CCA-adding enzyme (PDB id: 1UET), undergo sheet-to-helix transitions upon binding with the tRNA (PDB id: 2DRB) (Fig. 4D). Loop-to-helix transitions also contributes significantly to the change in accessibility (average $|\Delta A_P| = 34.3$ Å^2) at the non-interface regions, whereas, transitions from helix-to-loop or loop-to-sheet contribute moderately.

The effect of intermolecular H-bonds on accessibility. We evaluate the effect of intermolecular H-bonds on the change in solvent accessibility of amino acid residues and nucleotides at the protein-RNA interfaces. We find the change in accessibility is significant for the residues that are not involved in any H-bond with the partner nucleotides across the interfaces as compared to those involved in H-bond (Fig. 5A). This trend is observed in the entire dataset as well as among the different classes. The average $|\delta A_P|$ is 61.3 Å^2 for residues involved in H-bonds across the interface, whereas, those do not participate in H-bonds have an average of 93 Å^2.

At the RNA side, the change in accessibility is significantly higher for nucleotides that do not involved in any H-bond compared to those involved in H-bond across the interface (Fig. 5B). This phenomenon is observed in the entire dataset as well as among the different classes. Interestingly, different trend is observed in $|\delta A_R|$ among phosphate, sugar and bases. Among those involved in H-bonds across the interface, the highest change in average $|\delta A_R|$ is observed in bases (38.3 Å^2), followed by phosphate (32.5 Å^2) and sugar (14.4 Å^2). On the contrary, those do not participate in any H-bonds across the interface, the highest change in average $|\delta A_R|$ is observed in bases (183.7 Å^2), followed by sugar (163 Å^2) and phosphate (83.5 Å^2).
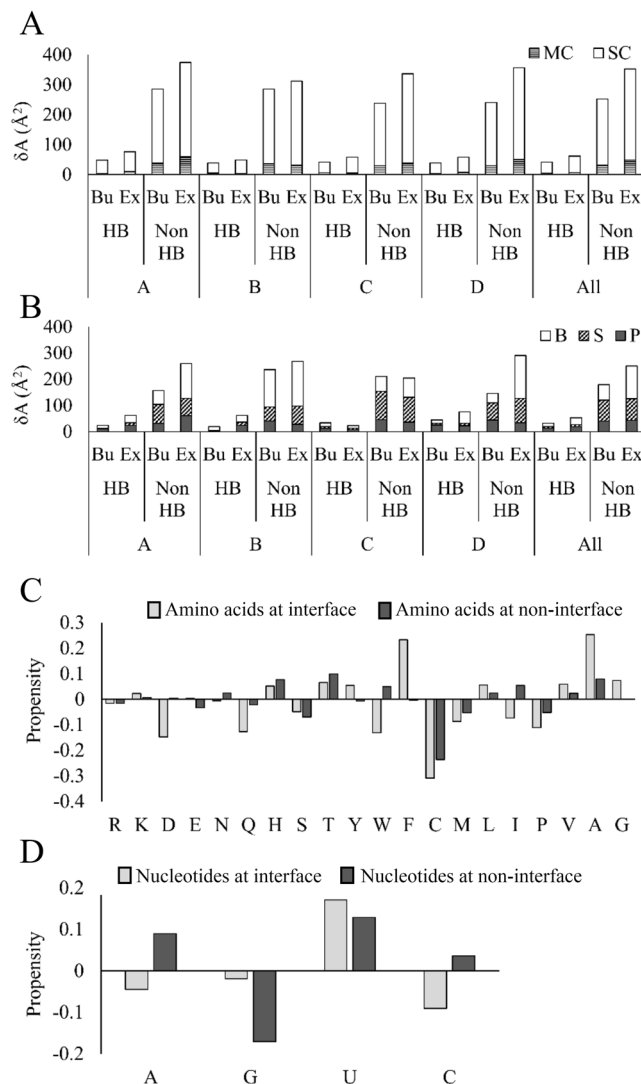
Figure 5. Distribution of δA in main chain and side chain calculated on 116 RBPs **(A)**, and in phosphate, sugar and bases calculated on 31 RNAs **(B)**. The average values are presented for Buried (Bu) and Exposed (Ex) surfaces of different class of complexes. Values for both H-bond (HB) and non-H-bond (Non HB) residues are given. Propensities of **(C)** amino acid residues and **(D)** nucleotides to get exposed or buried upon binding.

Accessibility of residues and nucleotides upon binding. The propensity of amino acid residues to get buried or exposed upon binding is shown in Fig. 5C. Upon binding, a positive propensity signifies that the residue prefers to get exposed while a negative propensity indicates their preference to get buried. Among the positively charged residues, Arg shows little preference to get buried both at the interface and at the non-interface regions, while, Lys shows opposite trend at both the regions. Among the negatively charged residues, Asp shows strong preference to get buried at the interface, while, Glu shows similar preference at the non-interface region, but with a lesser extent. Between Asn and Gln, the former shows preference to get exposed only at the non-interface, while the later shows preference to get buried both at the interface and at the non-interface regions. Among the neutral polar residues, His and Thr prefer to get exposed, whereas, Ser prefers to get buried both at the interface and at the non-interface regions. Among the three aromatic residues, Tyr and Phe both prefer to get exposed at the interface with a different magnitude, while Trp prefers to get buried at the interface and get exposed at the non-interface. Both the sulphur containing residues, Cys and Met, prefer to get buried both at the interface and the non-interface regions, however, with a different magnitude; the former have stronger preference than the later. Among the hydrophobic residues, Leu, Val and Ala prefer to get exposed both at the interface and the non-interface regions, while Gly prefers to get exposed only at the interface. On the contrary, Pro prefers to get buried both at the interface and the non-interface regions. Ile behave differently, it prefers to get buried at the interface and get exposed at the non-interface regions.

Among the four nucleotides, adenine and cytosine prefer to get buried at the interface and get exposed at the non-interface regions. Guanine prefers to get buried, while uracil prefers to get exposed both at the interface and at the non-interface regions (Fig. 5D).
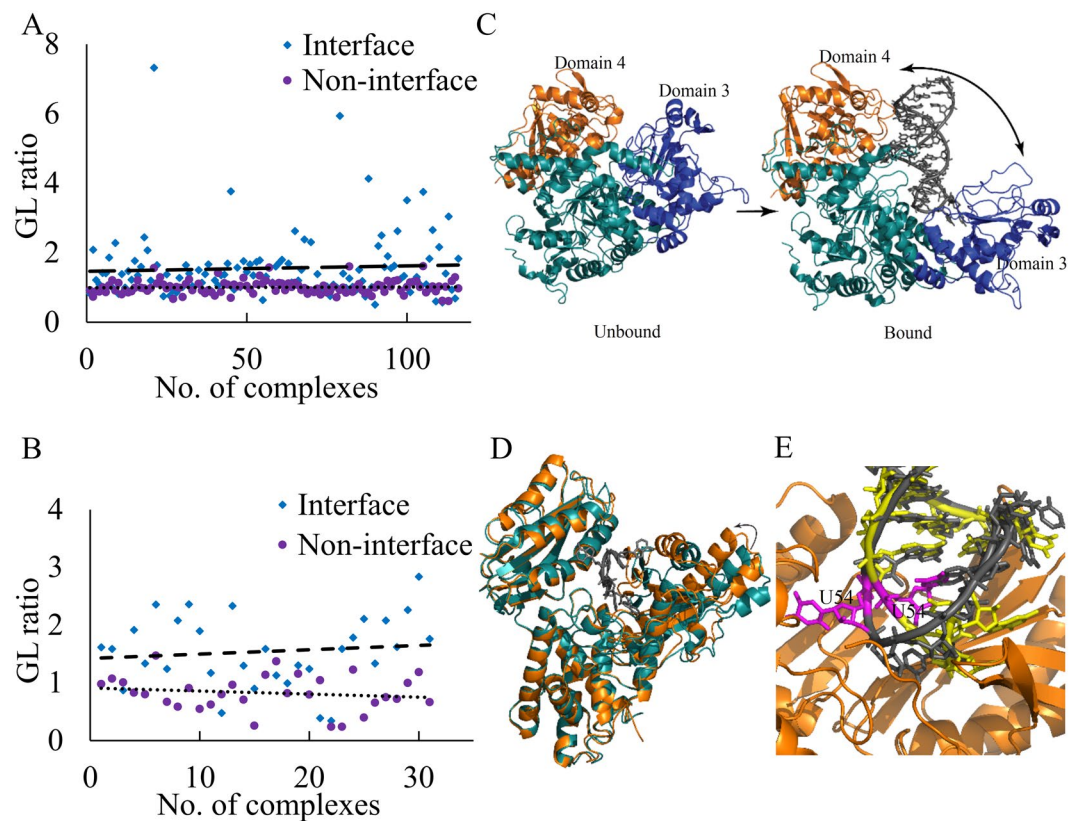
Figure 6. Gain or loss in accessibility. **(A)** The distribution of GL ratio of RBPs at interface and non-interface regions. **(B)** The distribution of GL ratio of RNAs at interface and non-interface regions. **(C)** In the unbound state of IRP1(PDB id: 2B3Y), domain 3 and 4 are in closed conformation, which transformed into open conformation upon binding with the RNA (PDB id: 3SNP). Both the domains move apart (bidirectional arrow), thereby increasing significant amount of surface to accommodate the RNA. Domain 3 and 4 are colored in blue and orange respectively, and the rest of the protein is colored in teal. **(D)** Example of "open-to-close" conformation change in poly(A) polymerase and oligo(A) RNA complex (PDB id: 2Q66). In the unbound state (color teal; PDB id: 2HHP), the binding cleft between N- and C-terminal domains remain wide open, which transformed into closed conformation upon binding with the RNA, hence losing the accessibility. **(E)** Superposed T-arm analogue RNA segment in bound (in grey; PDB id: 3BT7) and in unbound (in yellow; PDB id: 1EVV) states. The U54 (in magenta) in the unbound state remains inside the loop, which flips out to the active site upon binding with the 5-methyluridine methyltransferase TrmA (shown in orange).

Change in SASA can be used as a parameter to score protein-RNA decoys. Binding induced conformational transitions lead to change in SASA of individual atoms in interacting subunits. Few of the atoms gain accessible surface and few lose. We find the average gain to loss ratio of accessible surface area (GL ratio) upon binding is 1.7 and 1.0 (p-value = 1.6E-04, single tailed t-test) at the interface and at the non-interface regions, respectively. In majority of the cases, the ratio is close to one at the non-interface region. This ratio has never been used in any available protein-RNA docking algorithms[16], and may be efficiently use to score the flexible docking models to identify the near native solution. Figure 6A and 6B shows the distribution of the GL ratio in 115 RBPs and in 31 RNAs, respectively. The highest GL ratio (18.7) is found in the structure of iron regulatory protein 1 (IRP1) in complex with ferritin H IRE RNA (PDB id: 3SNP). This high ratio can be attributed to the large conformational change in IRP1 upon binding with the RNA, which is facilitated by a major rearrangement of the two domains of IRP1[17] (Fig. 6C), gaining 1279 Å² accessibility at the interface. The lowest GL ratio (0.5) is observed in complex between poly(A) polymerase and oligo(A) RNA (PDB id: 2Q66). In the polymerase, the catalytic site is located at the bottom of the cleft between N- and C-terminal domains of the polymerase[18]. In the unbound state, both the domains of the polymerase remain in open conformation and adopt closed conformation upon binding with the RNA, thereby losing 163.6 Å² surface area at the interface (Fig. 6D). The highest GL ratio (2.8) at the RNA binding surface is observed in the T-arm analogue RNA segment (PDB id: 1EVV) in complex with 5-methyluridine methyltransferase TrmA (PDB id: 3BT7). In the unbound state, U54 remains buried inside the T-loop of the tRNA and forms a reverse-Hoogsteen base pair with A58[19]. In the bound state, the loop changes its conformation and U54 flips out towards the active site of the enzyme, thereby gaining the surface accessibility of 310.4 Å² (Fig. 6E).

Discussion

Proteins and RNAs are flexible biomolecules that often undergo conformational changes upon binding. These conformational changes affect their solvent accessibility at the bound state as compared to their unbound states. Moreover, solvent accessibility can be directly attributed to the overall entropy of the system[5,6], and any change in accessibility upon complexation may contribute to the free energy of binding. Hence, it is utmost important to understand how the accessibility changes upon protein-RNA recognition. In this study, we have used the protein-RNA docking benchmark version 2[13] to evaluate the change in accessibility at the binding interfaces as well as at the non-interface regions of the partner molecules. The 3D structures used in this study are solved either by X-ray crystallography or by NMR, and are determined by different groups. Hence, it may be subject to serious artefacts. However, Fig. 2B shows that in most of the complexes, the non-interface regions of the proteins display only small changes in accessibility. The changes occur at the non-interface regions in both directions and their differences in unbound and bound is insignificant. Thus, the increased accessibility observed at the interface in the complexes is probably not influenced by the experimental artefacts. In this study, all the calculations were performed on the equivalent atoms of the bound and the corresponding unbound structures. About 2% of all the amino acid residues were discarded from the bound structures due to the mismatches in the alignment with their unbound form. This affects the average B (reported in Table 1), which is smaller compared to the values calculated on the benchmark dataset[13]. Mismatches in the alignment may arise due to the missing residues, which can be attributed to the disorder to order transitions of the partner molecules upon binding[1,20]. This is exemplified in Figure 3B, which illustrates the superposed structures of L25 protein in unbound (PDB id: 1B75) and in bound (PDB id: 1DFU) conformations with 5S rRNA. Residues 14 to 23 in L25 interact with the major groove of the RNA. This stretch of residues is unstructured in the unbound state, and changes to an α-helix in the bound state[15]. We have identified 10 interface residues in L25 that undergo loop-to-helix transitions upon binding with the rRNA with ΔA_P of 230 Å^2. Another example of disorder to ordered transition is found in the nuclear RNA exporter protein (TAP-NTD; PDB id: 3RW6) bound with the constitutive transport element (CTE) RNA (PDB id: 3RW7)[21]. Here, the linker of TAP-NTD, spanning from Ser194 to Leu204 (crystallographic data is missing from residue 199 to 202), is disordered in the unbound state. When bound to CET, the linker is well ordered and reorients the RRM and LRR domains in favorable binding conformation. This disorder to order transition accounts for a significant change in overall accessibility at the interface ($\delta A_P^{int} = -170$ Å^2).

Protein–RNA binding may results in four different possible changes in the solvent accessibility: the binding surfaces of both RBPs and RNAs get exposed or buried, RBPs get exposed but RNAs get buried, and vice-versa. Among the 21 $P_U R_U$ cases, we find in nine cases both the interacting surfaces get exposed upon binding, in five cases protein interfaces get exposed and RNA interfaces get buried, while in rest of the cases, RNA interfaces get exposed and protein interfaces get buried. This phenomenon can be correlated with the concept of induced fit in protein–RNA recognition[1]. We did not find any example where both protein and RNA interfaces get buried simultaneously upon binding. Negative δA indicates a gain in accessibility upon binding, which can be attributed to the exposure of molecular surfaces. On the other hand, a positive δA indicates loss in accessibility upon binding, which can be attributed to the burial of the molecular surfaces. The average δA_P (−120.5 Å^2, Table 1) suggests overall gain of solvent accessibility at the interface region of RBPs upon binding with RNAs, a phenomenon which is also observed in protein-protein recognition[8,22]. Except in two cases, the average δA_P and δA_R (Table 1) indicates that both RBPs and RNAs gain accessibility at the protein–RNA interfaces upon binding. Among two exceptions, in one case, the positive value of average δA_P in class B complexes can be attributed to the unstructured regions (Fig. 3B) along with the elongated N- and C-terminal loops in the majority of the ribosomal proteins, which are stabilized while forming the ribosomal assemblies[23]. In another case, the positive value of average δA_R at the interfaces in class C complexes can be supported by the molecular dynamics simulation study of double-stranded RNA, which explains that a stable A-form geometry of duplex RNA undergoes negligible changes while interacting with double-stranded RNA binding domains[24].

The interface regions undergo significant change in accessibility compared to the non-interface regions both in RBP and RNA. In RBPs, side chain atoms always contribute more to δA compared to main chain atoms. This can be attributed to the higher degrees of freedom of side chain atoms than the main chain. A moderate correlation (R = 0.6) between δA_P^{int} and i-rmsd[11] signifies that the side chain also play an important role in RNA recognition. The large movements of interacting domains in RBPs upon binding may or may not justify the large change in accessibility. Domains with large inter-domain buried area generally contribute to high δA_P^{int} that is correlated with high i-rmsd. On the other hand, the degree of change in accessibility does not always correlate with the i-rmsd if the interacting domains are connected with unstructured loop and the inter-domain buried surface area is small. In such cases, side chain conformations play the key role in change in accessibility. The change in accessibility can also be described in terms of binding free energy of the RBPs. Experimentally determined Gibbs free energy (ΔG), curated by Barik et al.[10], does not show any correlation with δA_P^{int} (we ignored the positive or negative sign of δA_P^{int} for this calculation). However, a moderate correlation (R = 0.5) is observed in 17 cases where the i-rmsd is greater than 1.0 Å. Moreover, the correlation slightly improves (R = 0.53) for eight cases where the i-rmsd is greater than 2.0 Å. This can be attributed to the collaborative effects of hydrophobic collapse and hydrophilic exposures of the interacting atoms that contribute in binding free energy. This observation can be supported with a study by Janin[25], which suggests that binding affinity of protein-protein interactions can be estimated by using only two geometric quantities: size of the interface and i-rmsd between bound and unbound forms of the interacting subunits. However, there are more accurate physics based empirical approaches to calculate ΔG, but the complexity of those algorithms make it computationally expensive and time consuming. At the interface, the extent of exposed surface of RBPs upon binding is always significantly higher than the buried surface. On the other hand, no such significant difference is observed at the non-interface region. Similar trend is

observed at the binding surface of the RNA; in 81% of cases, RNA gain accessibility upon binding. On the other hand, in majority of the cases (71%), non-interface region of the RNA lose accessibility upon binding.

The extent of gain and loss of solvent accessibility upon binding is almost equal at the non-interface region. On contrary, a significant change (either gain or loss) in solvent accessibility is observed at the interface region. This can be attributed to the exclusion of water molecules from the binding site[26]. Consequently, the hydrophobic residues get exposed and subsequent contacts between RBPs and RNAs make a stable conformation. Moreover, the predominant presence of positively charged and aromatic amino acid residues[27], having comparatively large and bulky side chains, provide higher degrees of flexibility at the interface compared to the non-interface region. Recently, many theoretical models have been developed to predict binding affinities and to discriminate the near native structures from the protein-protein decoys using MM-PBSA and MM-GBSA methods[28–30]. In both the approaches, the non-polar term directly correlates to the solvent accessible surface area or SASA. HawkRank, a scoring function developed by Feng *et al.*[31] describes a method to implement SASA-based solvation model to identify the near-native complexes from protein-protein decoys. In a similar context, GL-ratio can be used as a parameter to score the protein-RNA decoys generated by flexible docking algorithms. However, more detailed studies are required to optimize this parameter.

We observe frequent transitions of secondary structural elements in RBPs upon binding with its partner RNA. In the entire dataset, we find 308 transitions at the interface and 2,088 transitions at the non-interface regions. At the interface, loop-to-helix transitions are the most abundant followed by helix-to-loop, loop-to-sheet and sheet-to-loop. Loop-to-helix transition is frequently observed in RBPs, and RNA plays an important role to induce the folding[32]. Except the transitions from loop-to-sheet, all other transitions account for a negative change in ΔA_P, which signifies a relative gain in accessibility of interface residues upon binding (Supplementary Table S2). Interestingly, all the residues that retain their secondary structures at the interface gain accessibility upon binding. At the non-interface region, all the transitions, except helix-to-loop, account for a loss in accessibility. The average ΔA_P per residue is highest for sheet-to-helix transitions, which accounts for only 0.2% of all the transitions, followed by sheet-to-loop and loop-to-sheet transitions. Average ΔA_P for loop-to-helix transition is significantly low though it accounts for 32% of all the transitions at the non-interface region. The secondary structures of the non-interface regions, which remain unchanged upon binding, have significantly lower average ΔA_P compared to the interface regions (Supplementary Table S2). In the entire dataset, we do not observe any helix-to-sheet transitions both at interface and non-interface, and sheet-to-helix transitions at interface region.

Among the specific interactions, electrostatic forces between the positively charged amino acids and the negatively charged phosphate backbone of nucleotides have long range attractions, which "lure" the RBPs and their partner RNAs[27,33]. On the other hand, H-bonds are effective in short range contacts and play a crucial role in specific protein-RNA recognition[34]. In this study, we observe the interacting residues and nucleotides that are not involved in intermolecular H-bonds show significant change in accessibility upon binding compared to those involved in such interactions. Moreover, the positively charged amino acids, which are prevalent at the protein–RNA interfaces, display low propensity of gaining or losing accessibility. Uracil and guanine prefer to gain and lose accessibility, respectively both at the interface and the non-interface regions. This may be attributed to their preferential binding with the RBPs[27].

Materials and Methods

Dataset of protein-RNA complexes and their unbound structures.
Protein–RNA complexes and their unbound structures were taken from the protein–RNA docking benchmark version 2[13]. Modified residues and nucleotide bases in each structure were kept with their corresponding amino acids and bases by changing the keyword 'HETATM' to 'ATOM' in their atomic coordinate files taken from the Protein Data Bank (PDB)[35]. Each of the PDB file was cleaned following Barik *et al.*[36]. According to Bahadur *et al.*[34], the dataset was divided into four classes: (A) complexes with tRNA, (B) complexes with ribosomal proteins, (C) complexes with duplex RNA and (D) complexes with single-stranded RNA.

Calculation of solvent accessibility.
The size of the protein–RNA interface (B) was estimated by measuring SASA buried in contact. We calculated B using the following two equations:

$$B^B = A_P^B + A_R^B - A_{PR}^C \tag{1}$$

$$B^U = A_P^U + A_R^U - A_{PR}^C \tag{2}$$

where A_P, A_R and A_{PR} are the SASA of protein, RNA and protein-RNA complex, respectively. In the first equation, SASA values were calculated using the interacting partners taken from the complex, while in the second equation, they were calculated using their unbound structures. The difference in B upon binding is further calculated using the following equation:

$$\Delta B = B^U - B^B = \Delta A_P + \Delta A_R \tag{3}$$

where the corresponding change in accessibility of protein and RNA upon binding can be given by the following two equations:

$$\Delta A_P = A_P^U - A_P^B \tag{4}$$

$$\Delta A_R = A_R^U - A_R^B \tag{5}$$

Equations 4 and 5 quantify the effect of any conformational change on SASA upon binding in protein and RNA structures, respectively. We used the following equations to normalize the above values:

$$\delta A_{P/R}^{\text{int}} = \frac{\Delta A_{P/R}^{\text{int}}}{B_{P/R}^B} \times 1000 \text{ Å}^2 \tag{6}$$

$$\delta A_{P/R}^{non\text{-}\text{int}} = \frac{\Delta A_{P/R}^{non\text{-}\text{int}}}{A_{P/R}^B} \times 1000 \text{ Å}^2 \tag{7}$$

where $\delta A_{P/R}^{\text{int}}$ represent the normalized value at the protein or the RNA interface ($\Delta A_{P/R}^{\text{int}}$) and $\delta A_{P/R}^{non\text{-}\text{int}}$ represent the same at the non-interface region ($\Delta A_{P/R}^{non\text{-}\text{int}}$). SASA values were calculated using the program NACCESS[37], which implements the Lee and Richards[4] algorithm. Interface area was calculated using PRince[38]. Any atom belongs to amino acid residues or nucleotides is considered at the interface if SASA is lost upon binding. Non-interface regions are protein or RNA surfaces that are not included in the interface. All these calculations were performed over the equivalent atoms of the bound and their corresponding unbound structures of the interacting partners. The average values of SASA of interface (B^B) and non-interface (A^B) regions for each class of complexes are mentioned in Table 1. Pairwise alignment between the bound and the unbound structures of the interacting partners were carried out using ClustalW[39]. Secondary structures were assigned using the program DSSP[40]. According to Rost and Sander[41], α-helix, 3_{10}-helix and π-helix were categorized as helices (H), extended strand as strands (S) and isolated β-bridge, turn, bend and coils as loops (L). However, an isolated β-bridge (B) proceed by a coil (_) together considered as two strands (B_ = SS), whereas, two isolated β-bridge with a coil in between is considered as three loops (B_B = LLL).

The propensity of an amino acid residue or nucleotide to get exposed or buried upon binding was calculated using the following equation:

$$P_i = \ln \frac{f_i \Delta A^{\text{exp}}}{f_i \Delta A^{bur}} \tag{8}$$

where

$$f_i \Delta A^{\text{exp}} = \frac{\Delta A_i^{\text{exp}}}{\sum\limits_{i=1}^{n} \Delta A^{\text{exp}}} \text{ and } f_i \Delta A^{bur} = \frac{\Delta A_i^{bur}}{\sum\limits_{i=1}^{n} \Delta A^{bur}}$$

here, ΔA_i^{exp} and ΔA_i^{bur} are gain or loss in accessibility of amino acid or nucleotide of type i upon binding, respectively.

Data Availability. The data that support the findings of this study are available from the corresponding author RPB on request.

Conclusion

We evaluate the change in SASA of RBPs and their partner RNAs upon binding. We find in majority of the cases, the interface of both the binding partners gain accessibility upon binding. On the other hand, majority of RNA non-interface region lose accessibility. Interestingly, no such significant bias is observed at the non-interface region of RBPs. The change in accessibility is grossly attributed to the side chain conformation even though a moderate role of main chain conformation is also observed. Additionally, the significant change in accessibility is observed when the binding is more flexible including large domain movements that expose the inter-domain buried surfaces in RBPs. Besides, we also observed a significant change in accessibility associated with secondary structural transitions at the interface of RBPs upon binding. Close-to-open transitions upon binding lead to gain in accessibility, whereas, accessibility is lost in open-to-close transitions. The significant change in accessibility at the binding interface is an intrinsic feature of both the binding partners governed by the binding induced flexibility in protein-RNA recognition, and may have implications in designing flexible docking algorithms.

References

1. Williamson, J. R. Induced fit in RNA-protein recognition. *Nat. Struct. Mol. Biol.* **7**, 834–837, https://doi.org/10.1038/79575 (2000).
2. Leulliot, N. & Varani, G. Current topics in RNA−protein recognition: Control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**, 7947–7956, https://doi.org/10.1021/bi010680y (2001).
3. Calabretta, S. & Richard, S. Emerging roles of disordered sequences in RNA-binding proteins. *Trends Biochem. Sci.* **40**, 662–672, https://doi.org/10.1016/j.tibs.2015.08.012 (2015).
4. Lee, B. & Richards, F. M. Interpretation of protein structures - estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400, https://doi.org/10.1016/0022-2836(71)90324-X (1971).
5. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338–339, https://doi.org/10.1038/248338a0 (1974).
6. Chothia, C. Structural invariants in protein folding. *Nature* **254**, 304–308, https://doi.org/10.1038/254304a0 (1975).
7. Marsh, J. A. & Teichmann, S. A. Relative solvent accessible surface area predicts protein conformational changes upon binding. *Structure* **19**, 859–867, https://doi.org/10.1016/j.str.2011.03.010 (2011).

8. Chakravarty, D., Guharoy, M., Robert, C. H., Chakrabarti, P. & Janin, J. Reassessing buried surface areas in protein–protein complexes. *Protein Sci.* **22**, 1453–1457, https://doi.org/10.1002/pro.2330 (2013).

9. Marsh, J. A. Buried and accessible surface area control intrinsic protein flexibility. *J. Mol. Biol.* **425**, 3250–3263, https://doi.org/10.1016/j.jmb.2013.06.019 (2013).

10. Barik, A., Nithin, C., Karampudi, N. B., Mukherjee, S. & Bahadur, R. P. Probing binding hot spots at protein-RNA recognition sites. *Nucleic Acids Res.* **44**, e9, https://doi.org/10.1093/nar/gkv876 (2016).

11. Barik, A., C, N., P, M. & Bahadur, R. P. A protein-RNA docking benchmark (I): nonredundant cases. *Proteins: Struct. Funct. Bioinform.* **80**, 1866–1871, https://doi.org/10.1002/prot.24083 (2012).

12. Pérez-Cano, L., Jiménez-García, B. & Fernández-Recio, J. A protein-RNA docking benchmark (II): Extended set from experimental and homology modeling data. *Proteins: Struct. Funct. Bioinform.* **80**, 1872–1882, https://doi.org/10.1002/prot.24075 (2012).

13. Nithin, C., Mukherjee, S. & Bahadur, R. P. A non-redundant protein-RNA docking benchmark version 2.0. *Proteins: Struct. Funct. Bioinform.* **85**, 256–267, https://doi.org/10.1002/prot.25211 (2017).

14. Tu, C. *et al.* Structure of ERA in complex with the 3′ end of 16S rRNA: implications for ribosome biogenesis. *Proc. Natl. Acad. Sci. USA* **106**, 14843–14848, https://doi.org/10.1073/pnas.0904032106 (2009).

15. Lu, M. & Steitz, T. A. Structure of *Escherichia coli* ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-Å resolution. *Proc. Natl. Acad. Sci. USA* **97**, 2023–2028, https://doi.org/10.1073/pnas.97.5.2023 (2000).

16. Miao, Z. & Westhof, E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comp. Biol.* **11**, e1004639, https://doi.org/10.1371/journal.pcbi.1004639 (2015).

17. Walden, W. E. *et al.* Structure of dual function iron regulatory protein 1 complexed with Ferritin IRE-RNA. *Science* **314**, 1903–1908, https://doi.org/10.1126/science.1133116 (2006).

18. Balbo, P. B. & Bohm, A. Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis. *Structure* **15**, 1117–1131, https://doi.org/10.1016/j.str.2007.07.010 (2007).

19. Alian, A., Lee, T. T., Griner, S. L., Stroud, R. M. & Finer-Moore, J. Structure of a TrmA–RNA complex: A consensus RNA fold contributes to substrate selectivity and catalysis in m5U methyltransferases. *Proc. Natl. Acad. Sci. USA* **105**, 6876–6881, https://doi.org/10.1073/pnas.0802247105 (2008).

20. Varadi, M., Zsolyomi, F., Guharoy, M. & Tompa, P. Functional advantages of conserved intrinsic disorder in RNA-binding proteins. *PLOS ONE* **10**, e0139731, https://doi.org/10.1371/journal.pone.0139731 (2015).

21. Teplova, M., Wohlbold, L., Khin, N. W., Izaurralde, E. & Patel, D. J. Structure-function studies of nucleocytoplasmic transport of retroviral genomic RNA by mRNA export factor TAP. *Nat. Struct. Mol. Biol.* **18**, 990–998, https://doi.org/10.1038/nsmb.2094 (2011).

22. Chakravarty, D., Janin, J., Robert, C. H. & Chakrabarti, P. Changes in protein structure at the interface accompanying complex formation. *IUCrJ* **2**, 643–652, https://doi.org/10.1107/S2052252515015250 (2015).

23. Korobeinikova, A. V., Garber, M. B. & Gongadze, G. M. Ribosomal proteins: structure, function, and evolution. *Biochemistry* **77**, 562–574, https://doi.org/10.1134/s0006297912060028 (2012).

24. Acevedo, R., Evans, D., Penrod, K. A. & Showalter, S. A. Binding by TRBP-dsRBD2 does not induce bending of double-stranded RNA. *Biophys. J.* **110**, 2610–2617, https://doi.org/10.1016/j.bpj.2016.05.012 (2016).

25. Janin, J. A minimal model of protein–protein binding affinities. *Protein Sci.* **23**, 1813–1817, https://doi.org/10.1002/pro.2560 (2014).

26. Barik, A. & Bahadur, R. P. Hydration of protein-RNA recognition sites. *Nucleic Acids Res.* **42**, 10148–10160, https://doi.org/10.1093/nar/gku679 (2014).

27. Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein–RNA interactions: a structural analysis. *Nucleic Acids Res.* **29**, 943–954 (2001).

28. Gohlke, H. & Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie (International ed. in English)* **41**, 2644–2676, 10.1002/1521-3773(20020802)41:15<2644::aid-anie2644>.0.co;2-o (2002).

29. Chen, F. *et al.* Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. *Physical chemistry chemical physics: PCCP* **18**, 22129–22139, https://doi.org/10.1039/c6cp03670h (2016).

30. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *Journal of computational chemistry* **32**, 866–877, https://doi.org/10.1002/jcc.21666 (2011).

31. Feng, T. *et al.* HawkRank: a new scoring function for protein-protein docking based on weighted energy terms. *Journal of cheminformatics* **9**, 66, https://doi.org/10.1186/s13321-017-0254-7 (2017).

32. Bahadur, R. P., Kannan, S. & Zacharias, M. Binding of the bacteriophage P22 N-peptide to the boxB RNA motif studied by molecular dynamics simulations. *Biophys. J.* **97**, 3139–3149, https://doi.org/10.1016/j.bpj.2009.09.035 (2009).

33. Law, M. J. *et al.* The role of positively charged amino acids and electrostatic interactions in the complex of U1A protein and U1 hairpin II RNA. *Nucleic Acids Res.* **34**, 275–285, https://doi.org/10.1093/nar/gkj436 (2006).

34. Bahadur, R. P., Zacharias, M. & Janin, J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res.* **36**, 2705–2716, https://doi.org/10.1093/nar/gkn102 (2008).

35. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242, https://doi.org/10.1093/nar/28.1.235 (2000).

36. Barik, A. C, N., Pilla, S. P. & Bahadur, R. P. Molecular architecture of protein-RNA recognition sites. *J. Biomol. Struct. Dyn.* **33**, 2738–2751, https://doi.org/10.1080/07391102.2015.1004652 (2015).

37. Hubbard, S. J. & Thornton, J. M. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London* **2** (1993).

38. Barik, A., Mishra, A. & Bahadur, R. P. PRince: a web server for structural and physicochemical analysis of protein-RNA interface. *Nucleic Acids Res.* **40**, W440–444, https://doi.org/10.1093/nar/gks535 (2012).

39. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680, https://doi.org/10.1093/nar/22.22.4673 (1994).

40. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637, https://doi.org/10.1002/bip.360221211 (1983).

41. Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599, https://doi.org/10.1006/jmbi.1993.1413 (1993).

42. Xue, S., Calvin, K. & Li, H. RNA recognition and cleavage by a splicing endonuclease. *Science* **312**, 906–910, https://doi.org/10.1126/science.1126629 (2006).

43. Zhang, Y. & Li, H. Structure determination of a truncated dimeric splicing endonuclease in pseudo-face-centered space group P21212. *Acta Crystallographica Section D* **60**, 447–452, https://doi.org/10.1107/S0907444903029482 (2004).

44. Unge, J. *et al.* A mutant form of the ribosomal protein L1 reveals conformational flexibility. *FEBS Lett.* **411**, 53–59, https://doi.org/10.1016/S0014-5793(97)00611-X (1997).

45. Tishchenko, S. *et al.* Structure of the ribosomal protein L1-mRNA complex at 2.1 A resolution: common features of crystal packing of L1-RNA complexes. *Acta Crystallographica Section D* **62**, 1545–1554, https://doi.org/10.1107/S0907444906041655 (2006).

Acknowledgements

Author Contributions

R.P.B. conceived the study. S.M. performed the study. S.M. and R.P.B. both analyse the data and contributed in manuscript writing.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-28373-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.