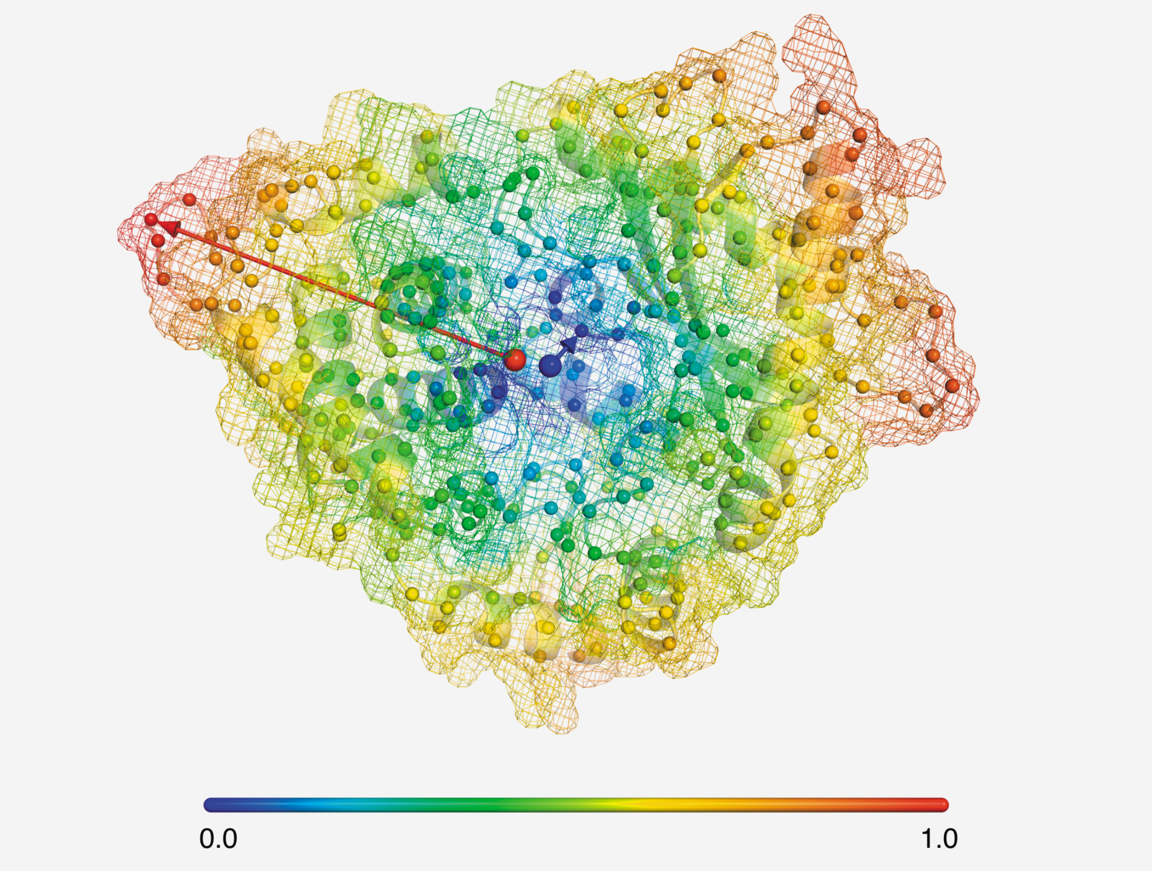
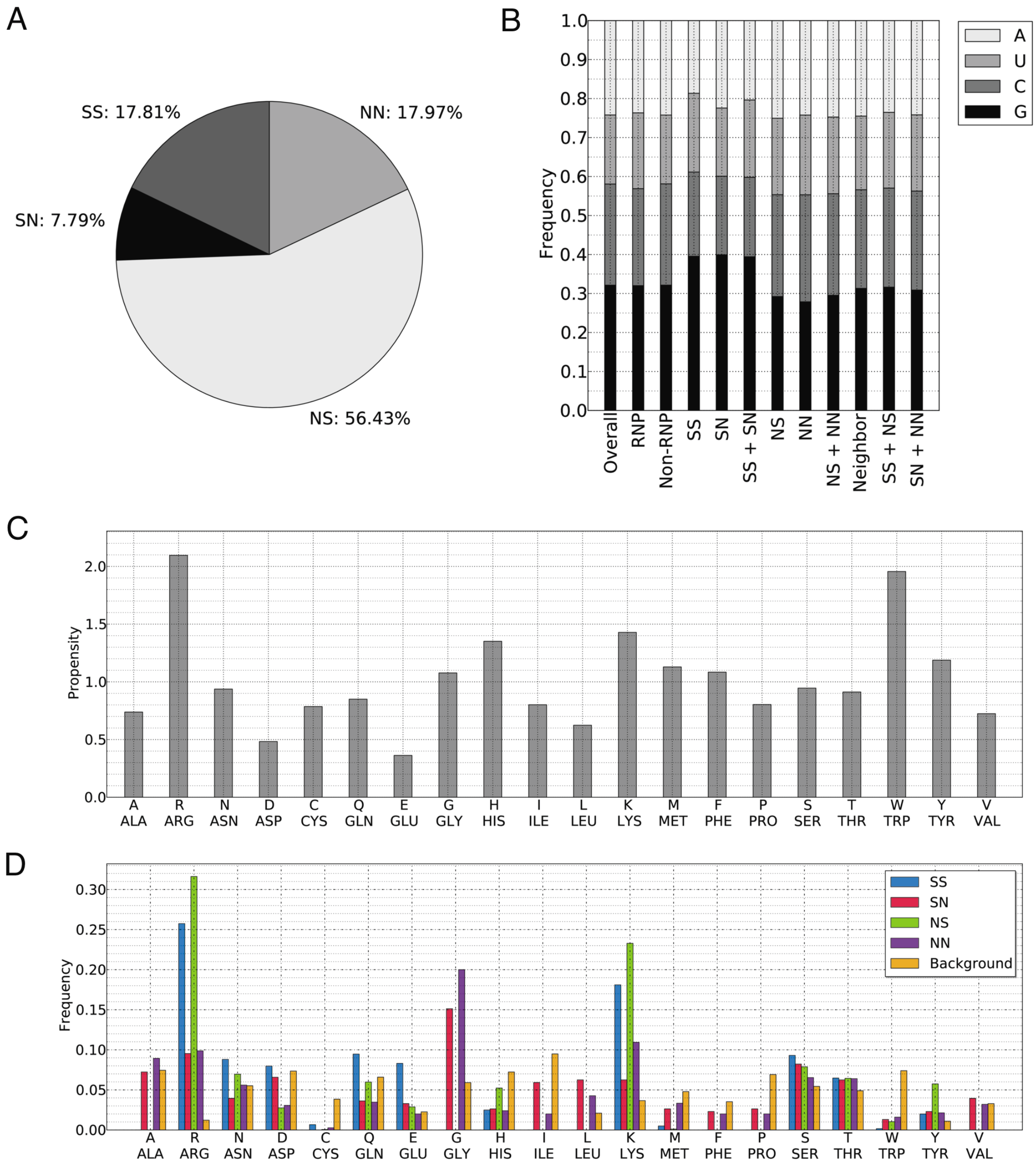
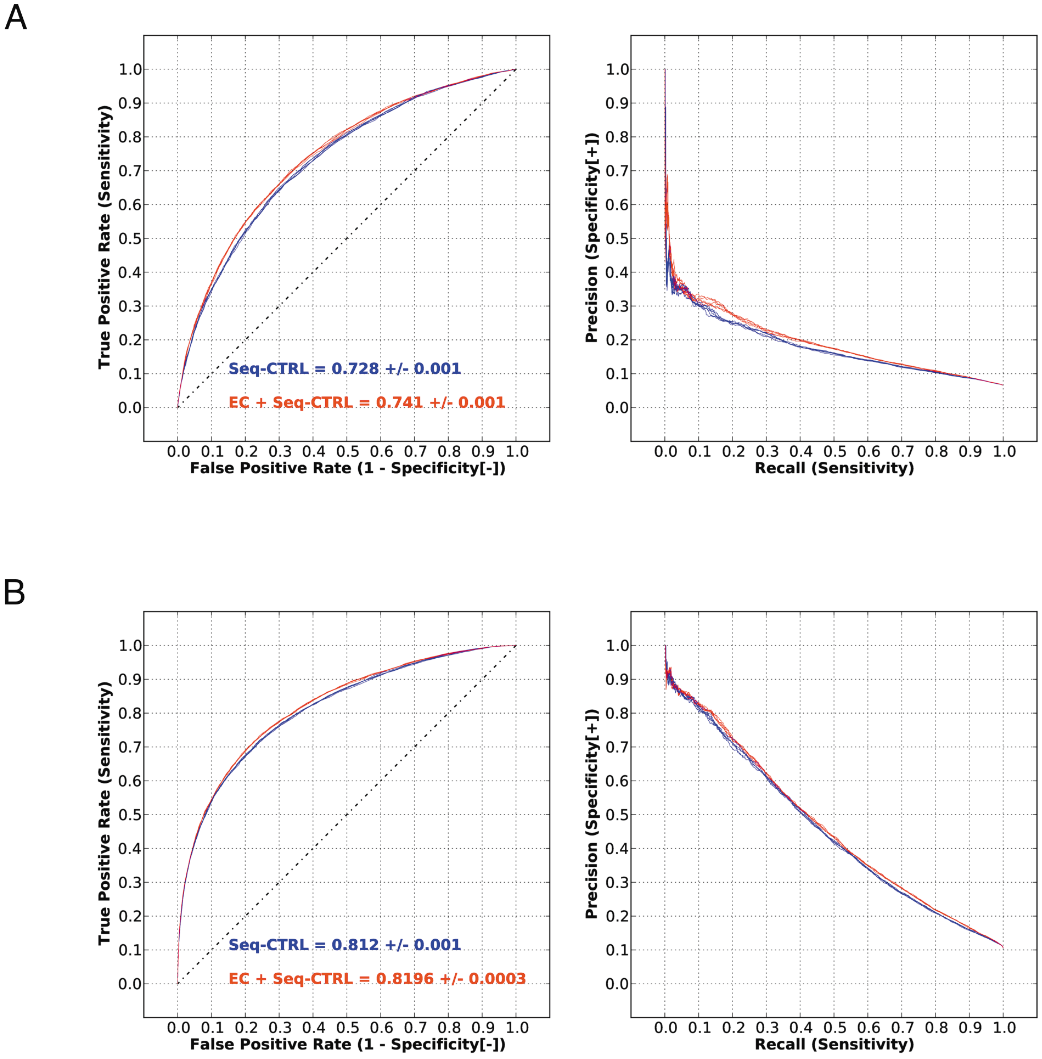
Supplementary Figures



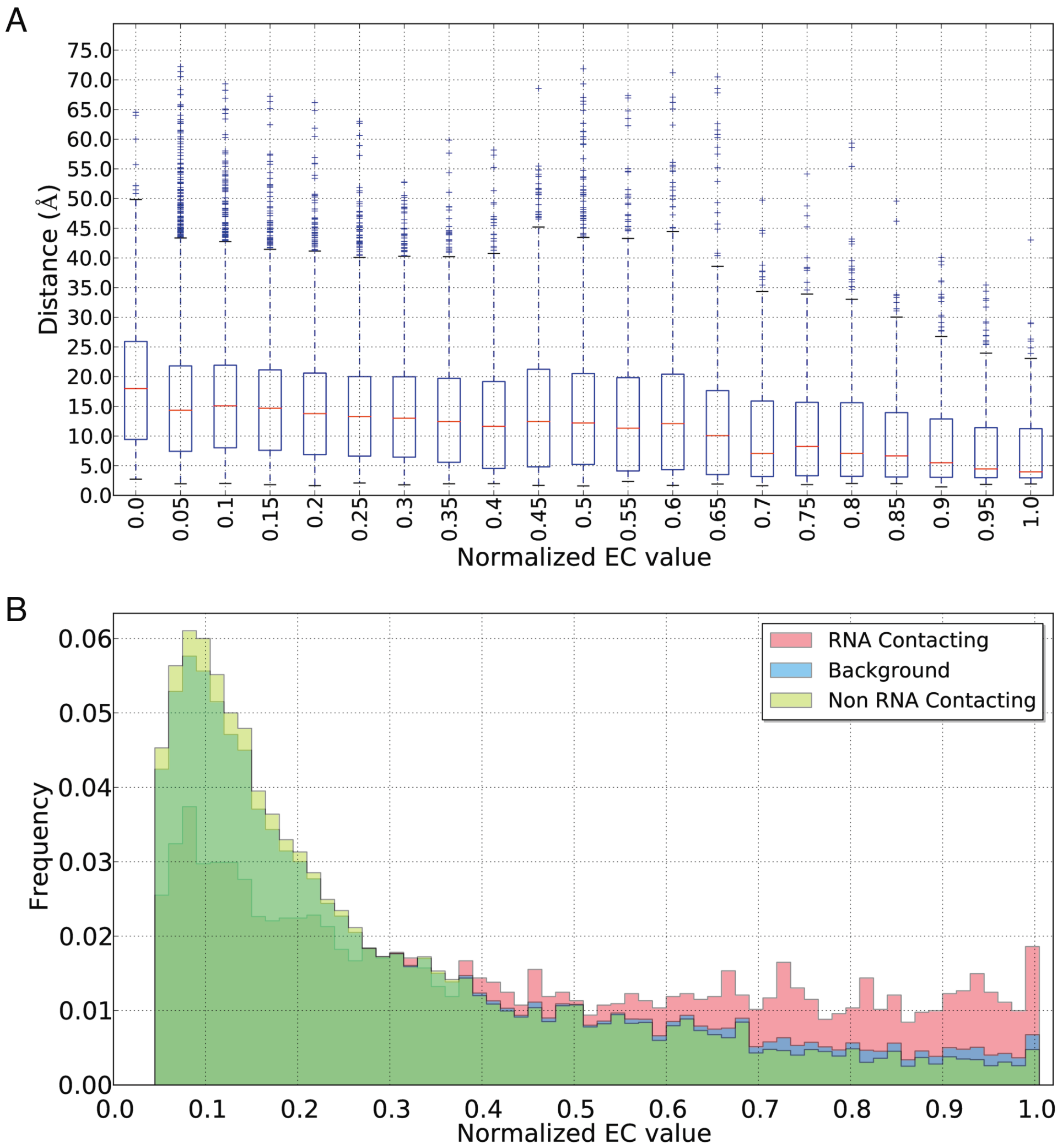
Supplementary figure 1. Geometrical features of residues reflected by Laplacain norms (LNs). The figure shows a mesh surface of RNA-binding protein 2'-5'-oligoadenylate synthase 1 (PDB entry: 4IG8) with high (low) values colored red (blue) according to Laplacian norms (rescaled from 0.0 to 1.0) calculated on a global scale. Carbon alpha atoms of residues are represented as small spheres and colored by their LNs. The arrow in red, pointing from the geometric center of the target residue's neighbors (large sphere in red) to the target residue, indicates the largest LN in this structure. In the same manner, the smallest LN of the structure is indicated by an arrow in blue. The LN value is proportional to target arrow length. When a surface residue is localized on a convex surface under a given scale, its LN will be larger. Conversely, a surface residue on a concave surface will have a smaller LN. For buried residues, the mean LN will be smaller than for exposed ones.



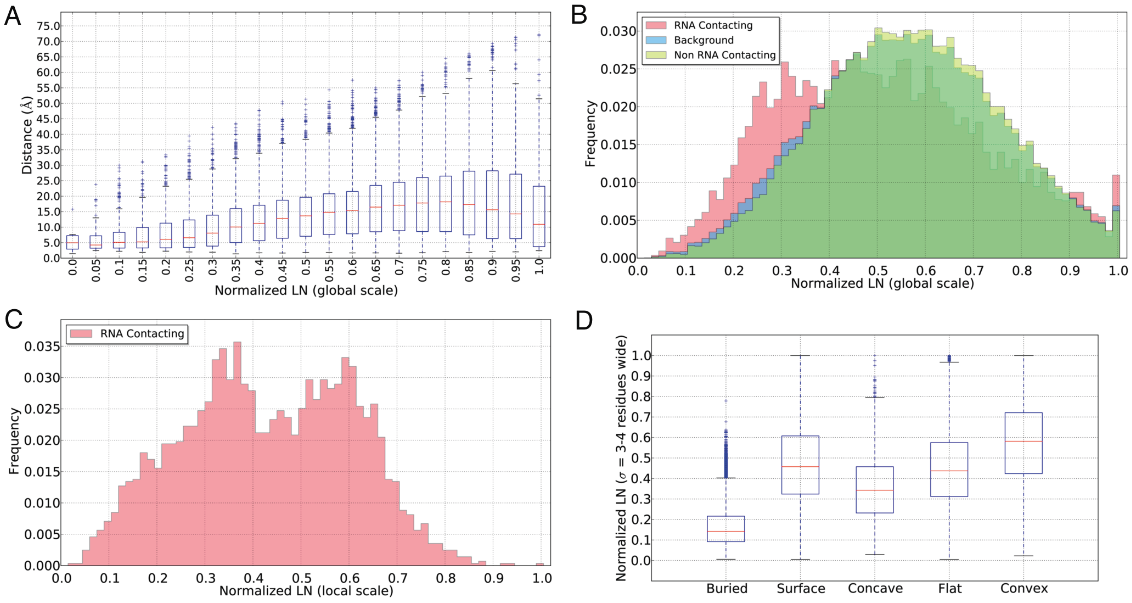
Supplementary figure 2. Contact preferences of protein-RNA complexes. A, the relative distribution of hydrogen bonds in different contact types. Contact pattern of protein and RNA can be either specific (S) or non-specific (N), considering side chain or backbone mediated contacts, respectively. The resulting four contact types can be defined as $\mathrm{S_{RNA}\_S_{PROT}}$ (SS), $\mathrm{S_{RNA}\_N_{PROT}}$ (SN), $\mathrm{N_{RNA}\_S_{PROT}}$ (NS), and $\mathrm{N_{RNA}\_N_{PROT}}$ (NN). Hydrogen bonds were identified for the full dataset using HBPLUS. RNA backbone and protein side chain are the most common type of hydrogen-bond (55.68 %). B, a histogram of nucleobase frequency in different contact types. The definition of types are as follows: Overall--nucleotides of RNA chains in the dataset; RNP--nucleotides of protein binding regions; Non-RNP--nucleotides of non-protein binding regions; SS + SN--nucleotides in contact via their side-chain; NS + NN--nucleotides in contact using their backbone; Neighbors--nucleotides that form no hydrogen bond, but stay in spatial proximity with the protein (< 4 Å); SS + NS--nucleotides in contact with protein side-chain; SN + NN--nucleotides in contact with the protein backbone. We can see that nucleobase composition of contacts mediated by side-chains (SS and SN) are enriched in G and U, and depleted in C, when compared with the overall composition. RNA backbone contacts (NS and NN), however, favor C and disfavor G. These differences are statistically significant. C, interface propensities of RNA-binding residues as calculated by comparing the interface ASA fraction with the surface ASA fraction of each amino acid type. When the propensities are greater than 1, the corresponding amino acid types are more likely to contact RNA when presented on protein surfaces. D, amino acid composition frequency for different contact types. Background: amino acid composition of all protein chains in the full dataset.



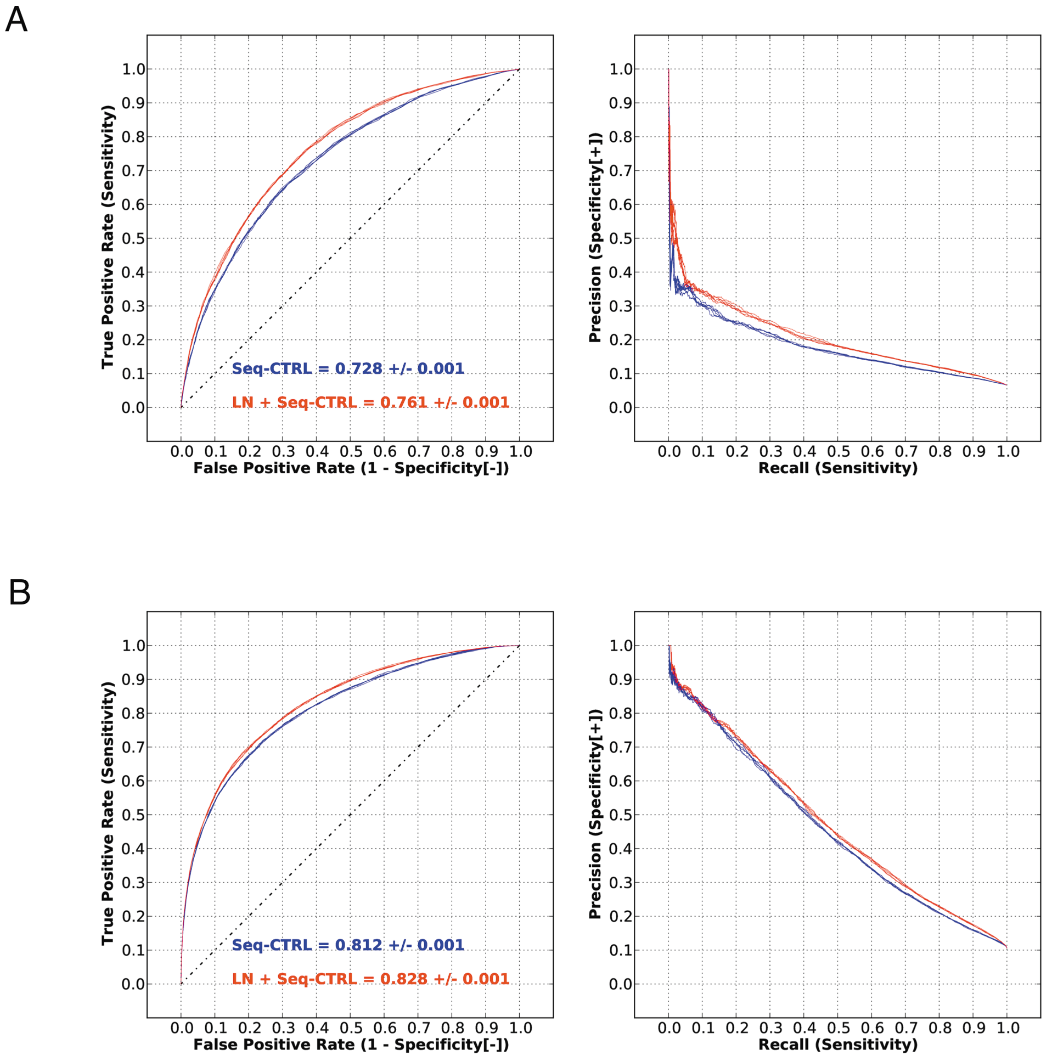
Supplementary figure 3. Performance of the network trained with the additional EC feature. Here, each line represents the ROC curve extracted from five-fold validation. Network training was repeated 5 times to check stability of non-ribosomal (A) and full (B) datasets. Blue curves indicate the performance of the control sequence-based method, with a mean value and standard deviation of AUCs calculated from five training repetitions, and red curves show the performance after introducing the EC feature. In each subfigure, the left panel shows the ROC curve, and the right panel shows the PR curve. Label “Seq-CTRL” reports the performance of the control method.



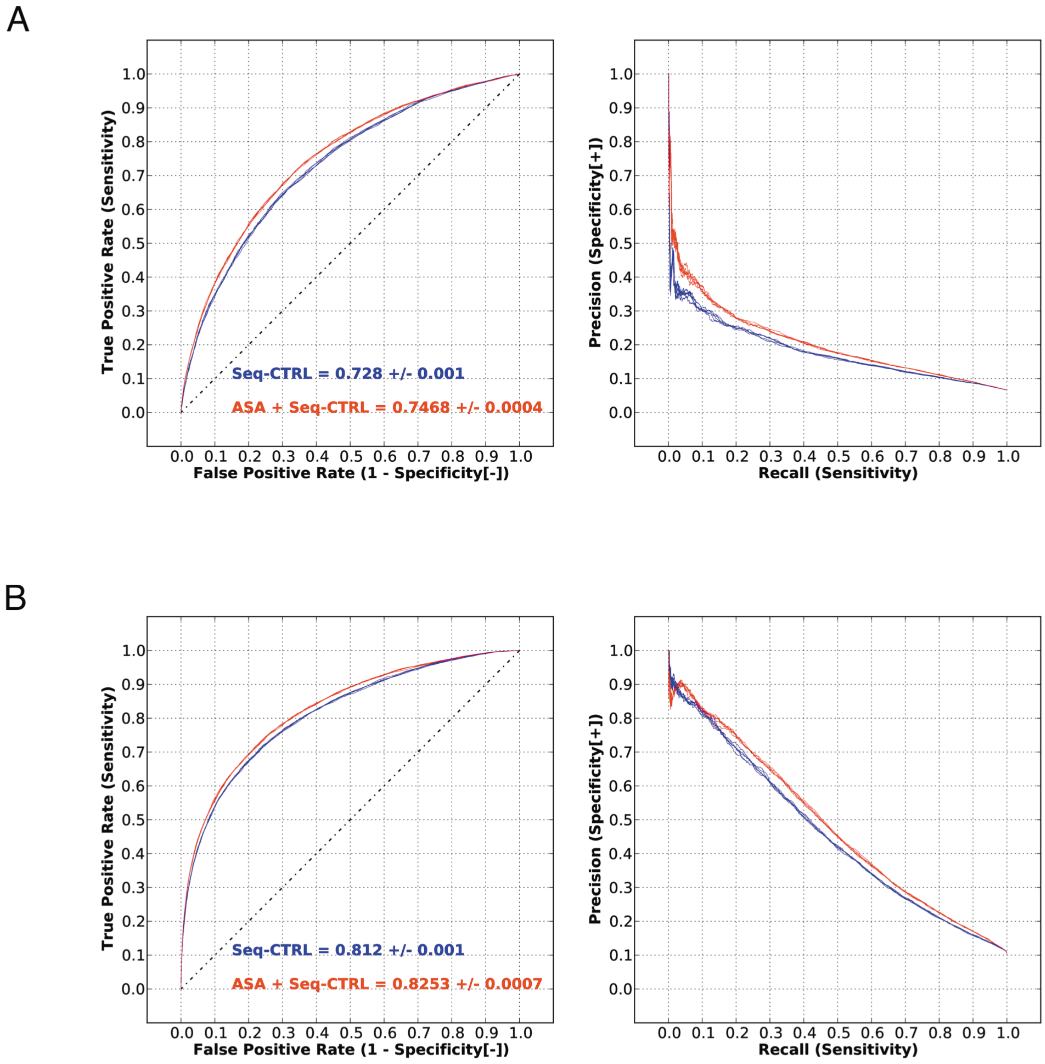
Supplementary figure 4. Comparison of evolutionary conservation (EC) of RNA binding and non-binding residues. Here we only took residues on the protein surface into consideration. Surface residues were defined as those have a ratio larger than 0.05 between the ASA calculated in the context of the whole protein and that calculated in isolation. In (A), a candlestick plot is used to illustrate the relationship between residue-RNA distances and residues' normalized EC values. The distance between a surface residue and RNA partner was defined as the minimal distance between any atom pair. Normalized EC values were binned with a bin size of 0.05. For each candlestick chart, box's lower and upper boundaries show 25% and 75% quartiles of binned data, respectively. And the median value is shown as a red line. Lower and upper whiskers indicate 1.5 times of interquartile range (IQR). Flier points are shown with “+” symbol in blue. We can see that the median value and deviation within the bin decrease, as the EC value increase. In (B), the distributions of normalized EC values of residues in contact with RNA and residues not in contact are compared. Label “Background” describes the distribution of all surface residues. We can see that for larger EC values, residues in contact with RNA are more enriched than background or non-contacting residues.



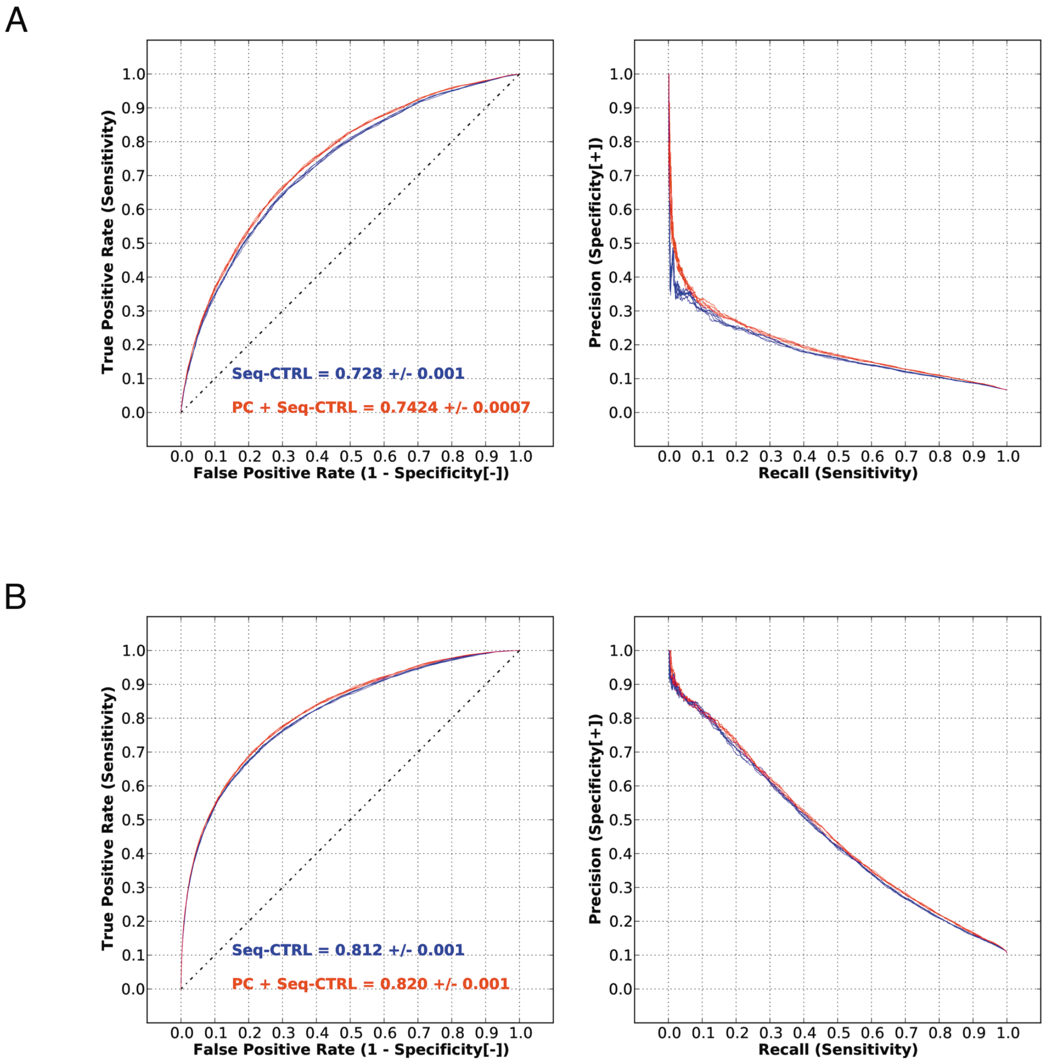
Supplementary figure 5. Laplacian norms (LNs) of RNA binding and non-binding residues. The distance between surface residues and RNA partner was calculated in the same way as in Suppl. Figure 4. Again, only surface residues were considered. Under a global scale, (A) shows the relationship between residue-RNA distances and residues' LN values, and (B) plots the distribution patterns of LN values from RNA binding/non-binding and all residues. In (A), the median value and the deviation of distances between surface residues and RNA increased with the normalized LN values (indicating a transition from concave to convex from left to right, respectively). However, as the LN values approached 1, the median distance decreased slightly. From the distribution patterns of LN values taken from RNA binding and non-binding residues, as well as all surface residues (B), it can be seen that RNA binding residues show a statistically significant (p-value < 1e-10) shift towards smaller LN values when compared with non-binding residues or background residues, which means that RNA is more likely to interact with residues located on globally concave surfaces. Interestingly, the frequency of RNA-binding residues with a LN value close to 1 is also higher. These residues are located at extremely convex points. In (C), we checked the distribution of local LN values for RNA-binding residues interacting with a globally concave surface (surface residues with global LN values smaller than 0.45 as shown in Suppl. Figure 5B). We can see that the distribution pattern shows two peaks; one exists at a relatively small local LN value corresponding to concave surfaces, while the other exists at moderately convex points. The frequency of contacts for flatter regions (i.e., around 0.5) is lower. In (D), we recalculated Laplacian norms using a sigma value spanning approximate 3-4 residues by taking the mean distance between C-alpha atom of the target residue and Cα atom of the third and the fourth adjacent neighbors. Then we plotted the distribution of residues' normalized LN values of 3-4 residues wide according to residues' locations in the 3D protein structures (buried, surface, concave, flat, and convex sites). Here, residues are considered on surface if they have a greater than 5% relative overall ASA (residues' ASA in protein divided by that in isolation). Otherwise, they are buried. Concave, flat, and convex residues, are picked out from surface residues and defined according to residues' global LN values. Concave residues are those, which have a smaller than 0.3 normalized global LN. Convex residues are those, which have a larger than 0.7 normalized global LN. And the rest residues are localized on more flatten surfaces. We can see that buried residues have a much smaller median LN values than surface residues. For concave, flat, and convex surface residues under a global scale, their median values of normalized LNs on a scale of 3-4 residues increase along with their global LN values. Large deviation of the normalized LNs under different categories shown in the figure is due to surface bumpiness of those globally defined surfaces. For example, a global concave surface could also be composed of many local concave, flat or convex surfaces within it. Therefore, roughness of a global surface will be more apparent when looking at more local LNs of surface residues.



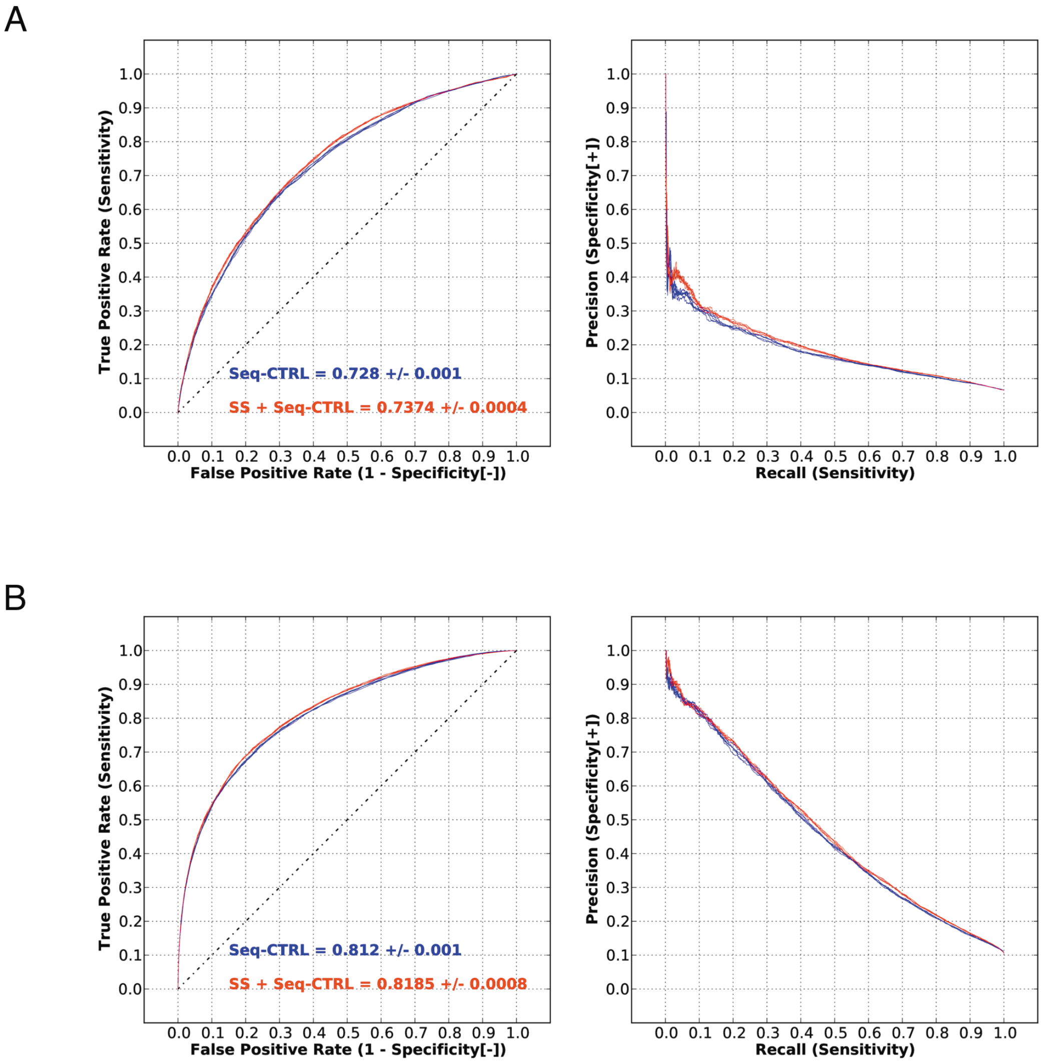
Supplementary figure 6. Performance of the network trained with the Laplacian norm (LN) feature in addition to the sequence features. Blue curves indicate the performance of the control sequence-based method, and red curves show the performance after introducing the LN feature.



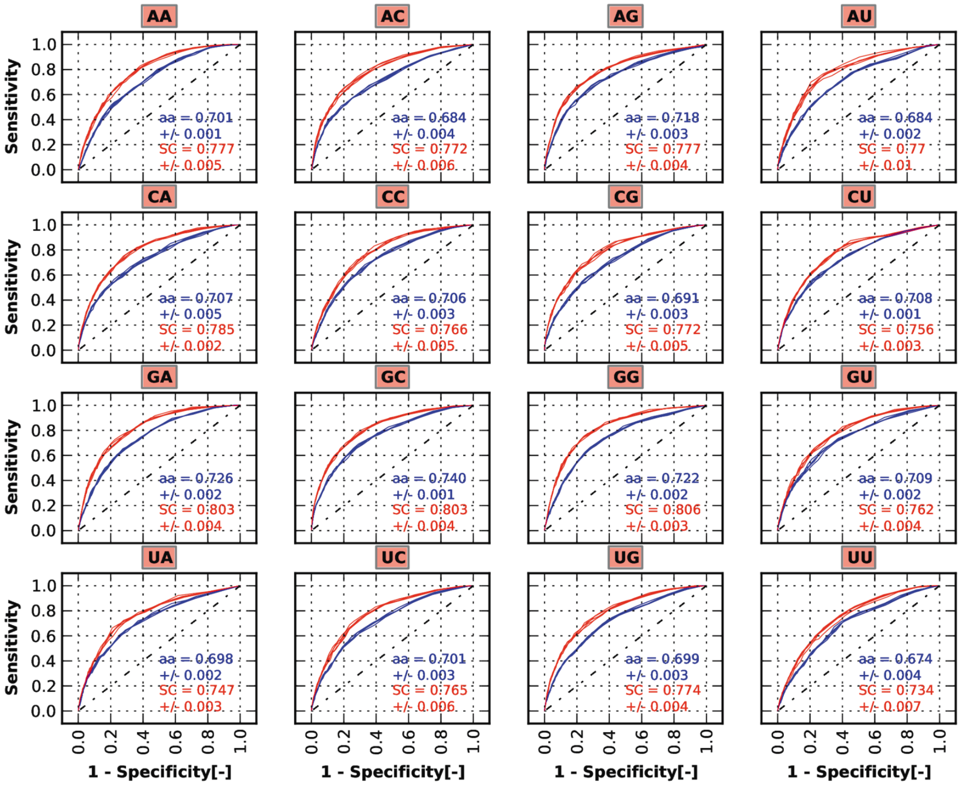
Supplementary figure 7. Performance of the network trained with the solvent assessable surface area (ASA) in addition to the sequence features. Blue curves indicate the performance of the control method and red curves show the performance after introducing the PC feature.



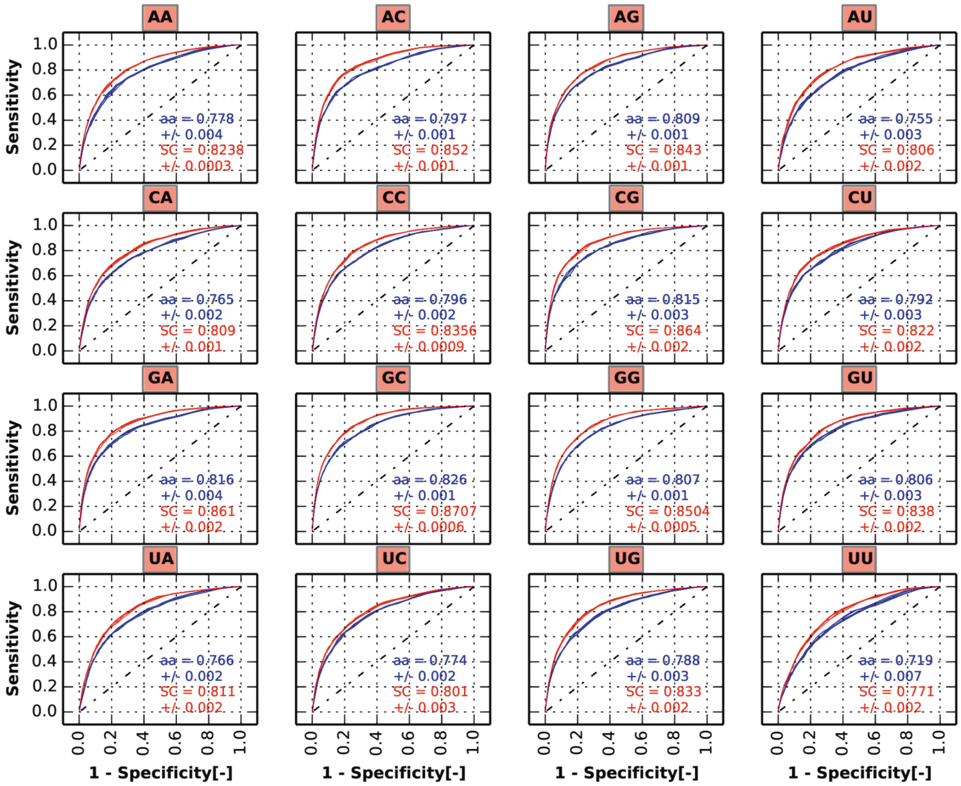
Supplementary figure 8. Performance of the network trained with the physicochemical (PC) features of neighboring residues in addition to the sequence features. Blue curves indicate the performance of the control method and red curves show the performance after introducing the PC feature.



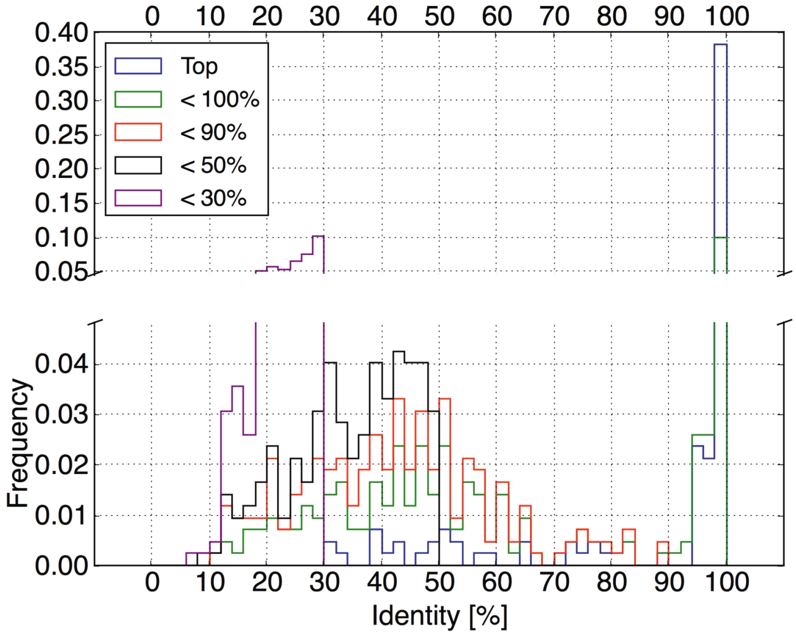
Supplementary figure 9. Performance of the network trained with the predicted secondary structure (SS) in addition to the sequence features. Blue curves indicate the performance of the control method and red curves show the performance after introducing the secondary structure feature.



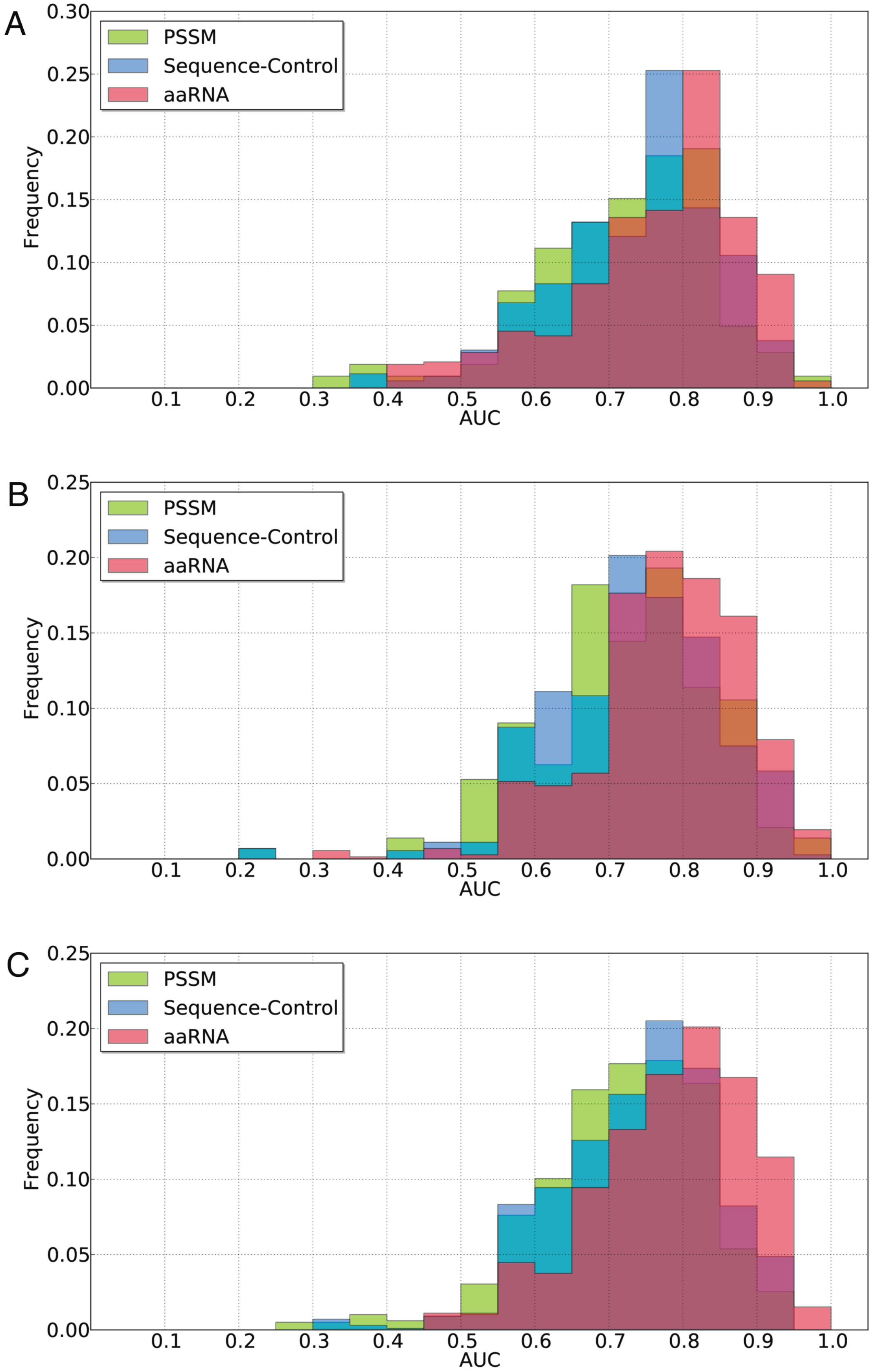
Supplementary Figure 10. Di-nucleotide prediction using all features on non-ribosomal set. For each type of di-nucleotide combination, a ROC curve is displayed. The label "SC" indicates the control method, and the label "aa" indicates the full predictor.



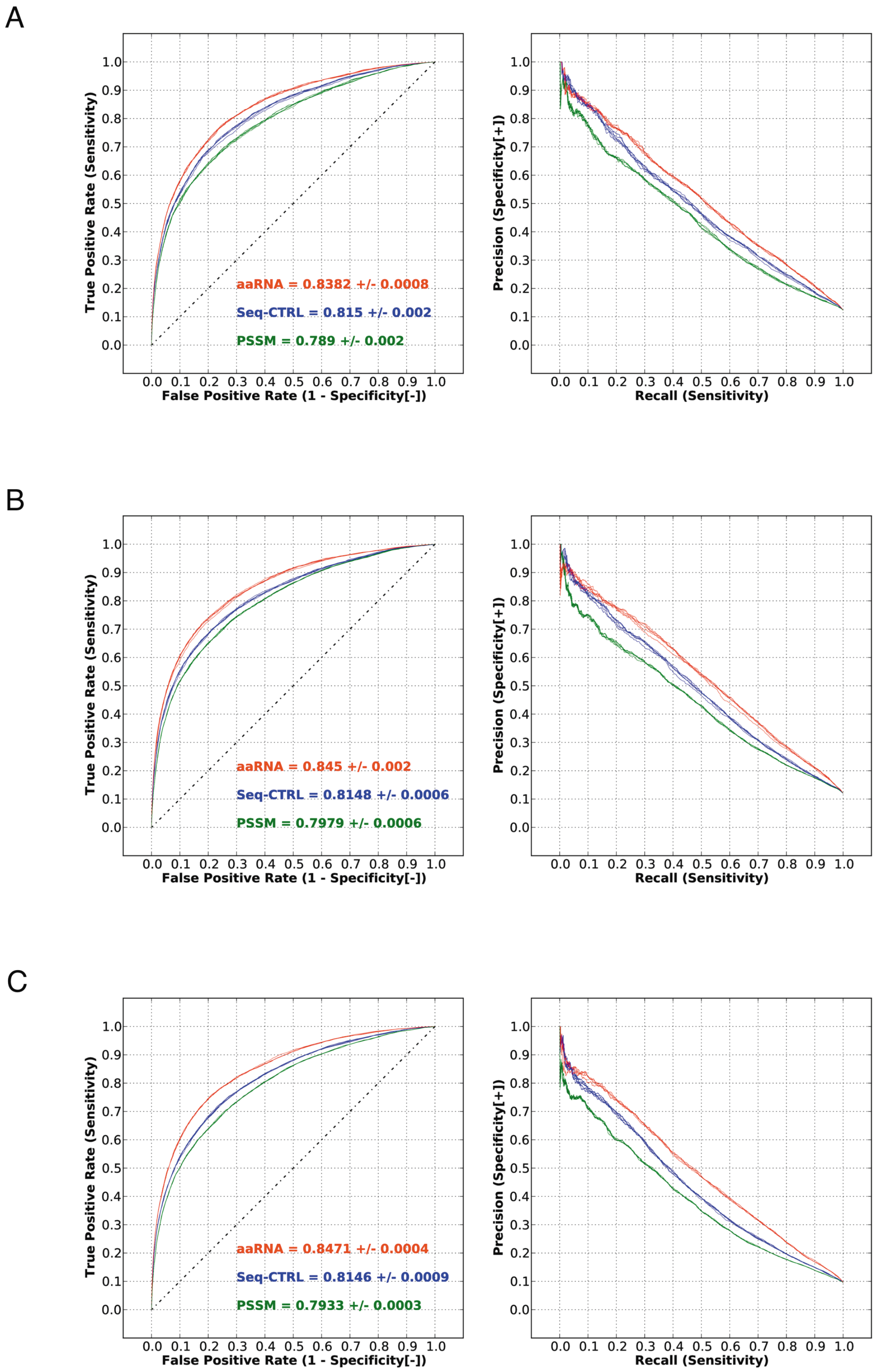
Supplementary Figure 11. Di-nucleotide prediction using all features on full set. For each type of di-nucleotide combination, a ROC curve is displayed. The label "SC" indicates the control method, and the label "aa" indicates the full predictor.



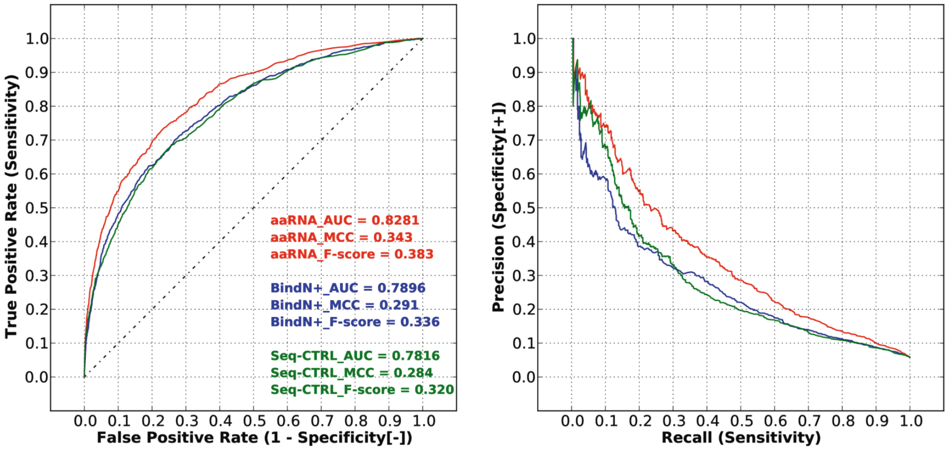
Supplementary figure 12. Identity distribution of template sequences under different thresholds. In the top group, best template was used while in the other groups the best template within the specified threshold (100%, 90%, 50%, and 30%) was used.


A
B
C
PSSM
Sequence-Control
aaRNA
Frequency
AUC
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

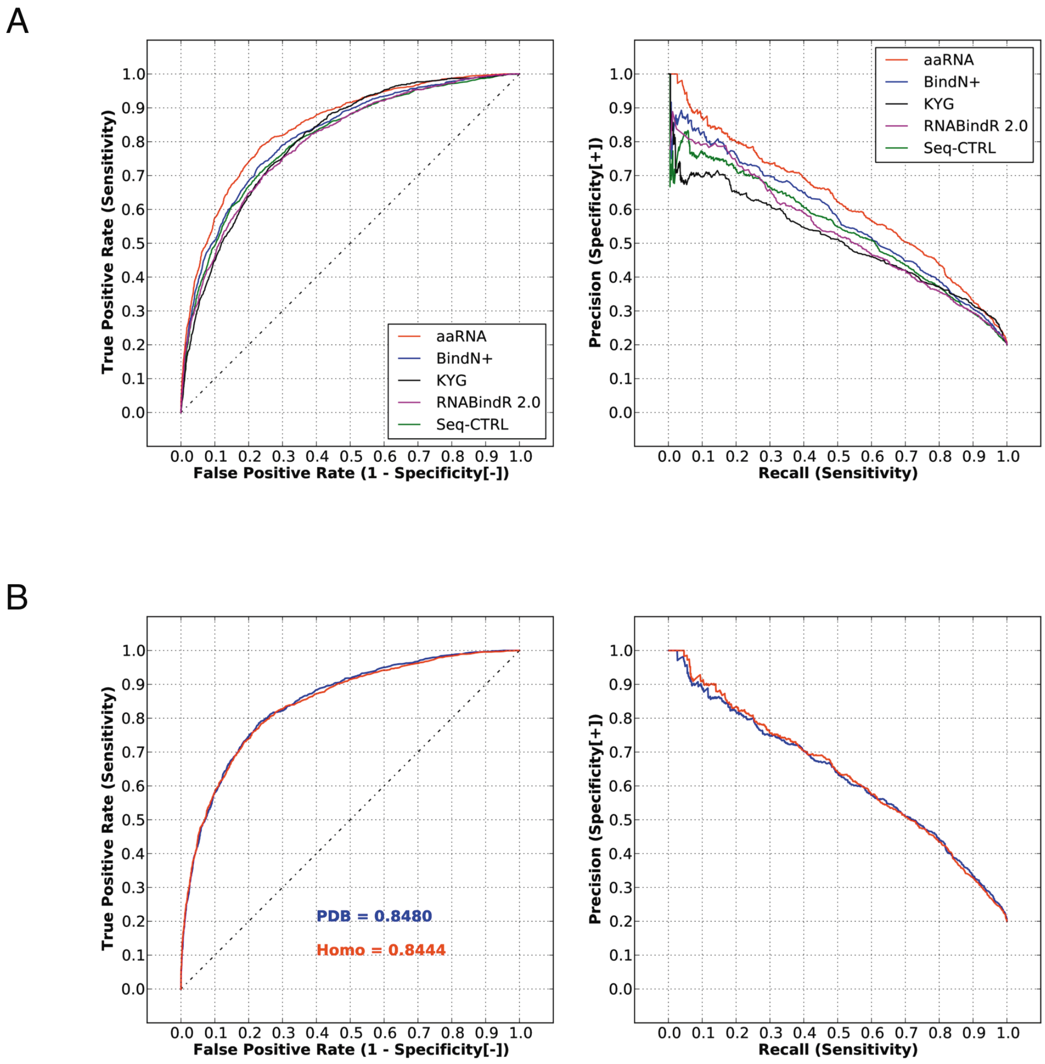
Supplementary figure 13. AUC distribution (protein-chain based) using different features and benchmarks RB106 (A), RB144 (B), and RB198 (C). In all three benchmarks, a larger AUC of aaRNA on average was statistically confirmed (See Table 3).


A
B
C
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity[-])
Precision (Specificity[+])
Recall (Sensitivity)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
aaRNA = 0.8382 +/- 0.0008
Seq-CTRL = 0.815 +/- 0.002
PSSM = 0.789 +/- 0.002
aaRNA = 0.845 +/- 0.002
Seq-CTRL = 0.8148 +/- 0.0006
PSSM = 0.7979 +/- 0.0006
aaRNA = 0.8471 +/- 0.0004
Seq-CTRL = 0.8146 +/- 0.0009
PSSM = 0.7933 +/- 0.0003

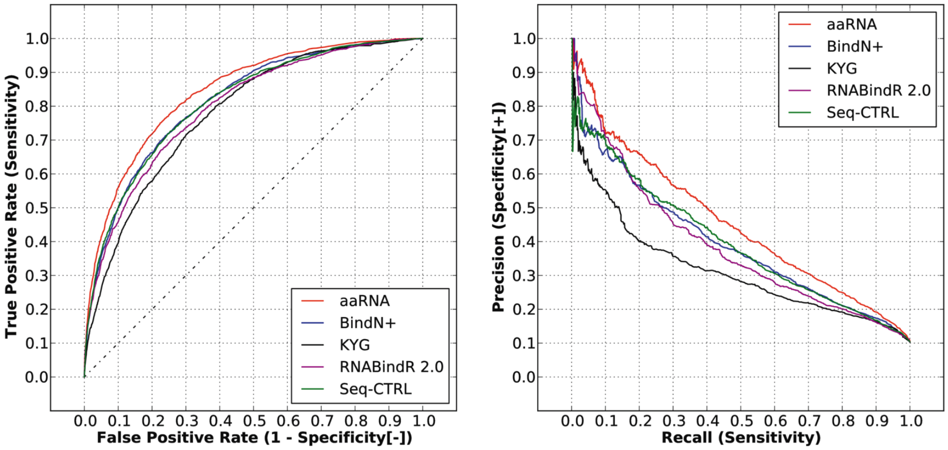
Supplementary figure 14. Performance of our feature-coding scheme on three benchmark datasets under a 3.5 Å distance cutoff for RNA-binding residues. The three benchmarks shows are RB106 (A), RB144 (B), and RB198 (C). The label "PSSM" indicates the AUC achieved with PSSM features only. The label "Seq-CTRL" indicates the result with the sequence-based control and the label "aaRNA" for all of our proposed features.



Supplementary figure 15. Prediction comparison of aaRNA and BindN+ methods on the dataset built from merged and redundancy-reduced RNABindR 2.0 datasets. In addition to the aaRNA and BindN+ methods, the label “Seq-CTRL” indicates the result with the sequence-based control method for reference.



Supplementary figure 16. Prediction on the RB44 dataset. A, Results using bound coordinates in the RB44 dataset. ROC and PR analysis was performed on the binding propensities returned by the predictors as shown in figure legends. B, aaRNA's performance (AUC) on homology models (Homo) built from the RB44 dataset sequences in comparison with bound conformations (PDB).



Supplementary figure 17. Prediction on the RB67 dataset. Comparison among ROC and PR curves of different predictors as listed in Table 5.

Supplementary Tables

Supplementary Table 1. The relative distribution of hydrogen bonds in different contact types.

Contact pattern of protein and RNA can be either specific (S) or non-specific (N), considering side chain or backbone mediated contacts, respectively. The resulting four contact types can be defined as SS, SN, NS, and NN. RNA and protein moieties participating in different contact types are listed. RNA backbone and protein side chain are the most common type of hydrogen-bond (56.43%)

Contact Type	RNA moiety	Protein moiety	Percentage (%)
SS	Side-chain	Side-chain	17.81
SN	Side-chain	Backbone	7.79
NS	Backbone	Side-chain	56.43
NN	Backbone	Backbone	17.97

Supplementary Table 2. The number of RNA binding/non-binding residues defined under different distance cutoff (3.5 and 5 Å) in three benchmark datasets (RB106, RB144, and RB198).

Benchmark	3.5 Å		5 Å	
	Binding	Non-binding	Binding	Non-binding
RB106	2,890	20,928	4,530	19,288
RB144	3,845	28,392	6,103	26,134
RB198	4,889	46,106	7,939	43,056

Supplementary Table 3. The number of homolog protein chains modeled when using templates of different sequence identities for RB141 and RB205 datasets, and homologies' averaged root-mean-square deviation (RMSD) from PDB structures.

Sequence Identity	Dataset		RMSD [Å]
	RB141 (# of chains)	RB205 (# of chains)	
Top	139	196	0.779
< 100%	127	182	1.353
< 90%	124	181	1.685
< 50%	121	177	1.825
< 30%	120	173	2.127