*Published online 14 February 2008*

*Nucleic Acids Research, 2008, Vol. 36, No. 5* ***e29***
*doi:10.1093/nar/gkn008*

# Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry

**Yao Chi Chen[1] and Carmay Lim[1,2,*]**

[1]Department of Chemistry, National Tsing Hua University, Hsinchu 300 and [2]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

Received November 13, 2007; Revised January 7, 2008; Accepted January 8, 2008

## ABSTRACT

**An RNA-binding protein places a surface helix, β-ribbon, or loop in an RNA helix groove and/or uses a cavity to accommodate unstacked bases. Hence, our strategy for predicting RNA-binding residues is based on detecting a surface patch and a disparate cleft. These were generated and scored according to the gas-phase electrostatic energy change upon mutating each residue to $Asp^-$/$Glu^-$ and each residue's relative conservation. The method requires as input the protein structure and sufficient homologous sequences to define each residue's relative conservation. It yields as output a priority list of surface patch residues followed by a backup list of surface cleft residues distant from the patch residues for experimental testing of RNA binding. Among the 69 structurally non-homologous proteins tested, 81% possess a RNA-binding site with at least 70% of the maximum number of true positives in randomly generated patches of the same size as the predicted site; only two proteins did not contain any true RNA-binding residues in both predicted regions. Regardless of the protein conformational changes upon RNA-binding, the prediction accuracies based on the RNA-free/bound protein structures were found to be comparable and their binding sites overlapped as long as there are no disordered RNA-binding regions in the free structure that are ordered in the corresponding RNA-bound protein structure.**

## INTRODUCTION

During post-transcriptional control, RNA metabolic processes such as splicing, polyadenylation, messenger RNA (mRNA) stability, mRNA localization and translation occur. All these chemical reactions involving RNA depend on the interactions between RNA and their target proteins. Consequently, identifying the key RNA recognition amino acid (aa) residues is important for understanding various critical biological processes such as mRNA processing, gene expression, protein synthesis, viral replication, cellular defense and developmental regulation (1). Despite the importance of protein–RNA interactions, they are less well understood compared with protein–DNA interactions primarily because RNA structures are more varied than DNA structures, resulting in a wider range of mechanisms for protein–RNA interactions. Whereas proteins seldom bind single-stranded (ss) DNA, they often bind ssRNA in a variety of secondary structures such as hairpins/stem-loops, bulges and loops (2). Furthermore, although RNA differs from DNA by the substitution of uracil for thymine and the presence of a 2′-OH group, their double-stranded (ds) conformations are quite different: dsRNA is found mainly in the A conformation with a narrow and deep major groove and a broad, shallow minor groove, but dsDNA is found mainly in the B conformation with a wide and accessible major groove but a narrow and deep minor groove (2).

Several statistical analyses of protein–RNA complex structures in the Protein Data Bank (PDB) (3) have revealed the following features of protein−RNA interactions (4–10). Proteins bind RNA either by placing a secondary structure such as an α-helix, a $3^{10}$-helix, a β-ribbon, or a loop in the groove of an RNA helix or by using β-sheet surfaces to create binding pockets in order to accommodate unstacked ssRNA bases (5). These empirical observations imply that the RNA binds either to a surface patch and/or to a cavity in the protein. The statistical analyses have also suggested some factors governing RNA-binding affinity and specificity. RNA-binding proteins achieve (i) RNA-binding affinity through favorable charge−charge interactions between positively charged Arg and Lys residues and the negatively charged RNA phosphate and (ii) specificity through directional hydrogen–bonding interactions and van der Waals (vdW) or non-polar contacts with specific bases

*To whom correspondence should be addressed. Tel: 011 886 2 2652 3031; Fax: 011 886 2 2788 7641; Email: carmay@gate.sinica.edu.tw

as well as steric exclusion of other bases (5,9,10). DNA-binding proteins employ a similar recognition strategy to bind dsDNA except that cavities and non-polar contacts are less frequently employed (5).

In addition to the aforementioned statistical analyses, a few studies have also attempted to predict RNA-binding proteins or residues. Given the protein sequence, support vector machines have been used to identify RNA-binding proteins and to assign them to different functional classes depending on the type of RNA bound (11–13). Support vector machines have also been used to predict RNA-binding sites from the protein sequence with ~69% accuracy, 70% specificity and 66% sensitivity (14). Other machine learning approaches using a neural network classifier (15) and a Naïve Bayes classifier (16), trained and tested on the same data set, yielded similar performance with an accuracy around 77%, specificity equal to 47% and sensitivity between 40–43%. In addition to sequence-based methods, residue and residue pairing preferences at the protein–RNA interface and the relative residue conservation at each position have been used to predict protein–RNA interface residues given the 3D structure and homologous sequences of a RNA-binding protein (17). However, the specificity$^+$ (the ratio of true positives to *predicted* interface residues) and sensitivity$^+$ (the ratio of true positives to *true* interface residues) of the prediction are anti-correlated; hence although the specificity$^+$ reached as high as 80%, the corresponding sensitivity$^+$ is only 10%. Interestingly, conservation alone was found to be a poor predictor of RNA interfaces, as the highest specificity$^+$ was only ~40%. This is because not all highly conserved surface residues constitute protein–RNA interface residues; thus considering conservation alone led to many false positives.

Here, given the 3D structure of a RNA-binding protein and its sequence homologs, we present a strategy for predicting RNA-binding regions on the basis of the following three criteria: The first criterion relies on the empirical observation that RNA-binding sites are comprised of multiple disparate regions, which are located not only on surface patches in analogy to DNA-binding sites, but also in binding pockets/cavities (5,10). The second criterion is founded on the *physical* principle that the RNA-binding site contains electropositive atoms providing charge–charge/dipole/quadrupole and hydrogen bonding interactions with electronegative RNA atoms. In the absence of RNA or water molecules, the positively charged or polar aa residues containing these electropositive atoms are in an unfavorable electrostatic environment (18); replacing one of these residues with a negatively charged $Asp^-/Glu^-$ would therefore alleviate the electrostatic repulsion among the electropositive atoms. The third criterion is based on an evolutionary principle that functionally important residues and aa residues in the vicinity, which form a cluster of spatially interacting residues, are usually highly conserved within the same family (19). Consequently, our strategy for predicting RNA-binding residues first generates irregular surface patches and clefts based on the second and third criteria. It then makes use of the first criterion to identify two disparate RNA-binding regions: (i) an irregular patch containing residues that are not only the most conserved, but also most electrostatically stabilized in the absence of solvent upon mutation to $Asp^-/Glu^-$ out of all the surface patches generated and (ii) a cleft containing the most conserved residues among the clefts generated. The method was tested first on a set of 69 structurally non-homologous proteins with RNA-bound structures and subsequently, on a smaller subset containing 18 proteins with 3D structures in the absence and presence of RNA.

## MATERIALS AND METHODS

### Test set of RNA-binding proteins

The method for predicting RNA-binding sites was tested on a heterogeneous set of 69 structurally non-homologous RNA-binding proteins with RNA-bound X-ray structures solved to $\leq 3$ Å resolution. This dataset was taken from our previous work (20), but five proteins with insufficient sequence homologs to define the conservation of each residue were omitted. The type of RNA bound to the protein such as dsRNA, mRNA, ribosomal RNA (rRNA), small nuclear RNA (snRNA), signal recognition particle RNA (srpRNA), transfer RNA (tRNA) and viral RNA (vRNA), the corresponding PDB code/chain and chain length are listed in the first three columns of Table 1, respectively.

The method for predicting RNA-binding sites was also tested on a heterogeneous set of 18 structurally non-homologous RNA-binding proteins with both RNA-bound and RNA-free X-ray structures solved to $\leq 3$ Å resolution (Table 2). This dataset was obtained by searching the PDB for all RNA-binding proteins with the same CATH code (21) as that of the representative RNA-bound protein and whose structures have been solved to $\leq 3$ Å resolution in the absence of RNA. Among the RNA-free protein structures with the same CATH code, the highest resolution structure was selected as the representative structure. The type of RNA bound to the protein, the PDB code/chain of the RNA-free and corresponding RNA-bound protein are listed in the first three columns of Table 2, respectively.

### Definition of true RNA-binding residues

A residue was defined as RNA-binding if any of its non-hydrogen atoms are within vdW contact or hydrogen bonding distance to any RNA non-hydrogen atom directly or indirectly via a bridging water molecule. The HBPLUS (22) program was used to compute all possible vdW contacts and hydrogen bonds in the protein–RNA complex structure, which are defined by a donor atom to an acceptor atom distance of 3.9 and 3.35 Å, respectively. The number of true RNA-binding residues in each protein–RNA complex structure, $n_T$, is listed in column four of Table 1.

### Definition of solvent accessible residues

The percentage aa accessibility is defined as the percent ratio of the solvent-accessible surface area (SASA) of the side-chain X in the protein to the SASA of X in the

**Table 1.** RNA-binding residue predictions based on the 3D structures of 69 structurally non-homologous protein–RNA complexes

| RNA type | | PDB-chain[a] | Chain length[b] | $n_T^c$ | Predicted patch/ Patch + Cleft[d] | | | | Predicted cleft[d] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $n_P^e$ | $n_{TP}^f$ | $n_{TP}^f/n_{max}^g$ | RPV[h] | $n_P^e$ | $n_{TP}^f$ | $n_{TP}^f/n_{max}^g$ | *RPV*[h] |
| dsRNA | 1. | 1di2-A[i] | 69 | 13 | 14 | 8 | 1 | 0.04 | 6 | 1 | 0.25 | 0.49 |
| | 2. | 1yz9-A | 220 | 25 | 10 | 0 | 0 | 1 | 9 | 0 | 0 | 1 |
| mRNA | 3. | 1av6-A | 289 | 16 | 11 | 5 | 0.83 | 0.04 | 29 | 8 | 1.14 | 0 |
| | 4. | 1fxl-A | 167 | 36 | 10 | 8 | 0.89 | 0.05 | 2 | 0 | 0 | 1 |
| | 5. | 1gtf-L[i,j] | 70 | 11 | 19 | 3 | 0.33 | 0.49 | – | – | – | – |
| | 6. | 1m8x-A | 341 | 33 | 24 | 2 | 0.40 | 0.82 | 23 | 11 | 2.75 | 0 |
| | 7. | 1wpu-A | 147 | 20 | 12 | 6 | 1 | 0.13 | 5 | 0 | 0 | 1 |
| | 8. | 1wsu-A[i] | 124 | 11 | 34 | 11 | 1 | 0.08 | 4 | 0 | 0 | 1 |
| | 9. | 2a8v-B[i] | 118 | 10 | 19 | 2 | 0.33 | 0.50 | 6 | 0 | 0 | 1 |
| rRNA | 10. | 1dfu-P[i] | 94 | 20 | 24 | 13 | 1 | 0.01 | 3 | 1 | 0.33 | 0.43 |
| | 11. | 1feu-A | 185 | 22 | 18 | 10 | 1 | 0.03 | 12 | 0 | 0 | 1 |
| | 12. | 1fjg-C | 206 | 35 | 25 | 16 | 0.89 | 0.12 | 11 | 5 | 0.56 | 0.31 |
| | 13. | 1fjg-D | 208 | 59 | 12 | 8 | 0.89 | 0.19 | 7 | 6 | 0.86 | 0.09 |
| | 14. | 1fjg-G | 155 | 31 | 11 | 1 | 0.09 | 0.68 | 8 | 0 | 0 | 1 |
| | 15. | 1fjg-I[i] | 127 | 52 | 27 | 26 | 1 | 0.02 | 20 | 15 | 0.75 | 0.21 |
| | 16. | 1fjg-J[i] | 98 | 30 | 13 | 12 | 1.09 | 0 | 4 | 3 | 0.75 | 0.21 |
| | 17. | 1fjg-K[i] | 119 | 31 | 17 | 8 | 0.67 | 0.28 | 23 | 14 | 1 | 0.03 |
| | 18. | 1fjg-M[i] | 125 | 46 | 27 | 20 | 0.95 | 0.14 | 24 | 17 | 0.85 | 0.19 |
| | 19. | 1fjg-N | 60 | 30 | 10 | 7 | 0.78 | 0.32 | 6 | 4 | 0.80 | 0.38 |
| | 20. | 1fjg-P | 83 | 45 | 17 | 14 | 0.82 | 0.28 | 7 | 7 | 1 | 0.18 |
| | 21. | 1fjg-S[i] | 84 | 23 | 22 | 17 | 1.21 | 0 | 4 | 0 | 0 | 1 |
| | 22. | 1fjg-T | 99 | 44 | 12 | 6 | 0.55 | 0.59 | 5 | 4 | 0.80 | 0.34 |
| | 23. | 1g1x-A[i] | 98 | 14 | 26 | 2 | 0.22 | 0.93 | 4 | 0 | 0 | 1 |
| | 24. | 1g1x-B | 88 | 29 | 12 | 5 | 0.56 | 0.50 | 13 | 8 | 0.89 | 0.20 |
| | 25. | 1g1x-H | 48 | 13 | 5 | 4 | 0.80 | 0.25 | 2 | 2 | 1 | 0.29 |
| | 26. | 1i6u-A | 127 | 21 | 11 | 8 | 1 | 0.01 | 10 | 7 | 1 | 0.02 |
| | 27. | 1mms-A | 133 | 29 | 10 | 1 | 0.12 | 0.72 | 10 | 5 | 0.62 | 0.24 |
| | 28. | 1mzp-A[i] | 213 | 31 | 20 | 4 | 0.50 | 0.46 | 23 | 14 | 1.75 | 0 |
| | 29. | 1sds-C | 112 | 22 | 11 | 5 | 0.56 | 0.24 | 8 | 8 | 1.14 | 0 |
| | 30. | 1vq8-1[i,j] | 56 | 51 | 18 | 17 | 0.94 | 0.79 | – | – | – | – |
| | 31. | 1vq8-3 | 92 | 60 | 8 | 8 | 1 | 0.10 | 3 | 2 | 0.67 | 0.71 |
| | 32. | 1vq8-A[i] | 237 | 123 | 72 | 71 | 1 | 0.17 | 20 | 18 | 0.90 | 0.40 |
| | 33. | 1vq8-B | 337 | 147 | 33 | 31 | 0.94 | 0.09 | 49 | 41 | 0.87 | 0.11 |
| | 34. | 1vq8-C | 246 | 117 | 10 | 10 | 1 | 0.14 | 4 | 2 | 0.50 | 0.61 |
| | 35. | 1vq8-D | 140 | 54 | 15 | 10 | 0.71 | 0.29 | 13 | 5 | 0.38 | 0.61 |
| | 36. | 1vq8-E[i] | 172 | 39 | 24 | 12 | 0.86 | 0.18 | 6 | 1 | 0.17 | 0.60 |
| | 37. | 1vq8-H | 160 | 52 | 19 | 5 | 0.29 | 0.82 | 19 | 12 | 0.70 | 0.24 |
| | 38. | 1vq8-J | 142 | 53 | 14 | 11 | 0.85 | 0.21 | 10 | 10 | 1 | 0.05 |
| | 39. | 1vq8-K | 132 | 37 | 9 | 1 | 0.11 | 0.73 | 11 | 0 | 0 | 1 |
| | 40. | 1vq8-L | 145 | 72 | 44 | 40 | 0.97 | 0.26 | 5 | 5 | 1 | 0.34 |
| | 41. | 1vq8-M | 194 | 120 | 13 | 12 | 0.92 | 0.25 | 5 | 4 | 0.80 | 0.59 |
| | 42. | 1vq8-N[i] | 186 | 74 | 44 | 39 | 1.03 | 0 | 13 | 13 | 1 | 0.05 |
| | 43. | 1vq8-O[i] | 115 | 42 | 12 | 9 | 0.82 | 0.05 | 10 | 5 | 0.56 | 0.53 |
| | 44. | 1vq8-P[i] | 143 | 81 | 17 | 17 | 1 | 0.03 | 7 | 6 | 0.86 | 0.25 |
| | 45. | 1vq8-Q | 95 | 57 | 17 | 15 | 0.88 | 0.15 | 5 | 4 | 0.80 | 0.57 |
| | 46. | 1vq8-R | 150 | 67 | 12 | 10 | 0.83 | 0.27 | 6 | 4 | 0.67 | 0.47 |
| | 47. | 1vq8-T | 119 | 55 | 17 | 13 | 0.87 | 0.13 | 5 | 3 | 0.60 | 0.50 |
| | 48. | 1vq8-U[i] | 53 | 15 | 11 | 7 | 1 | 0.11 | 5 | 3 | 0.75 | 0.23 |
| | 49. | 1vq8-V | 65 | 18 | 11 | 6 | 0.75 | 0.29 | 6 | 3 | 0.60 | 0.40 |
| | 50. | 1vq8-W[i] | 154 | 56 | 12 | 11 | 1 | 0.14 | 21 | 12 | 0.63 | 0.45 |
| | 51. | 1vq8-X[i,j] | 82 | 33 | 22 | 15 | 0.83 | 0.22 | – | – | – | – |
| SnRNA | 52. | 1ec6-B | 84 | 22 | 14 | 8 | 0.80 | 0.12 | 2 | 0 | 0 | 1 |
| | 53. | 1m8v-B[i,j] | 71 | 14 | 22 | 6 | 0.60 | 0.41 | – | – | – | – |
| | 54. | 1ooa-A[i] | 313 | 21 | 55 | 17 | 1 | 0.01 | 5 | 0 | 0 | 1 |
| SrpRNA | 55. | 1hq1-A | 76 | 18 | 10 | 5 | 0.62 | 0.29 | 7 | 0 | 0 | 1 |
| | 56. | 1jid-A | 114 | 22 | 17 | 10 | 1 | 0.07 | 10 | 6 | 0.75 | 0.11 |
| tRNA | 57. | 1b23-P | 405 | 29 | 10 | 5 | 0.62 | 0.05 | 28 | 9 | 0.90 | 0.03 |
| | 58. | 1c0a-A | 585 | 67 | 16 | 9 | 0.69 | 0.05 | 53 | 21 | 1 | 0.01 |
| | 59. | 1f7u-A | 606 | 76 | 17 | 2 | 0.17 | 0.53 | 12 | 12 | 1.33 | 0 |
| | 60. | 1gax-A | 862 | 58 | 11 | 8 | 0.80 | 0.01 | 69 | 8 | 0.36 | 0.39 |
| | 61. | 1h3e-A | 427 | 26 | 14 | 0 | 0 | 1 | 16 | 1 | 0.12 | 0.43 |
| | 62. | 1h4s-A[i] | 473 | 9 | 46 | 0 | 0 | 1 | 11 | 0 | 0 | 1 |
| | 63. | 1j1u-A[i] | 299 | 16 | 36 | 2 | 0.25 | 0.39 | 16 | 0 | 0 | 1 |
| | 64. | 1n78-A | 468 | 69 | 14 | 5 | 0.45 | 0.27 | 27 | 9 | 0.64 | 0.24 |

(Continued)

**Table 1.** Continued.

| RNA type | PDB-chain[a] | Chain length[b] | $n_T^c$ | Predicted patch/ Patch + Cleft[d] | | | | Predicted cleft[d] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $n_P^e$ | $n_{TP}^f$ | $n_{TP}^f/n_{max}^g$ | RPV[h] | $n_P^e$ | $n_{TP}^f$ | $n_{TP}^f/n_{max}^g$ | *RPV*[h] |
| | 65. 1q2r-A | 376 | 33 | 10 | 3 | 0.43 | 0.21 | 31 | 19 | 1.27 | 0 |
| | 66. 1qf6-A | 641 | 52 | 11 | 7 | 0.78 | 0.03 | 9 | 3 | 0.38 | 0.17 |
| | 67. 2fmt-A | 314 | 33 | 13 | 8 | 0.80 | 0.04 | 60 | 25 | 1.67 | 0 |
| vRNA | 68. 1ddl-B | 188 | 10 | 12 | 8 | 0.89 | 0.03 | 7 | 0 | 0 | 1 |
| | 69. 2bu1-A | 129 | 12 | 11 | 4 | 0.80 | 0.12 | 5 | 0 | 0 | 1 |

[a]PDB entry of the RNA-bound protein structure followed by the protein chain.
[b]The number of aa residues in the protein chain.
[c]The number of true RNA-binding residues based on the protein/RNA complex structure.
[d]The patch or cleft is generated and scored as described in the Materials and Methods section.
[e]The number of *predicted* RNA-binding residues, which is equal to the number of solvent accessible residues in the patch/cleft.
[f]The number of *true positives* or correctly predicted RNA-binding residues in the patch/cleft.
[g]The maximum number of true-positive RNA-binding residues among all the randomly generated patches. Proteins with $n_{TP}/n_{max} < 0.7$ in both predicted regions are highlighted by the gray background.
[h]The random pick value (RPV), is the fraction of *random* patches with true-positive RNA-binding residues $\geq n_{TP}$.
[i]The top-ranked patch is merged with nearby top-ranking clefts.
[j]For these small proteins, two disparate RNA-binding sites could not be found, and only a single RNA-binding site was predicted.

**Table 2.** RNA-binding residue predictions based on the 3D structures of 18 structurally non-homologous RNA-free and RNA-bound proteins[a]

| RNA type | PDB-chain | | RMSD[d] (Å) | $(n_{TP}/n^{max})_{patch}$ | | $(n_{TP}/n^{max})_{cleft}$ | | $f_{overlap}$[g] |
|---|---|---|---|---|---|---|---|---|
| | Free[b] | Bound[c] | | Free[e] | Bound[f] | Free[e] | Bound[f] | |
| mRNA | 1. 1a8v-A | 2a8v-B[h] | 0.94 | 0 | 0.33 | 1.25 | 0 | 0.68 |
| | 2. 1ib2-A | 1m8x-A | 1.12 | 0 | 0.40 | 1.60 | 2.75 | 0.20 |
| | 3. 1qaw-A[h] | 1gtf-L[h] | 0.28 | 0.25 | 0.33 | –[i] | –[i] | 0.74 |
| | 4. 1v39 | 1av6-A | 0.54 | 0.80 | 0.83 | 1 | 1.14 | 0.89 |
| | 5. 1wpv-A | 1wpu-A | 0.23 | 0.83 | 1 | 0 | 0 | 1.00 |
| rRNA | 6. 1ris[h] | 1g1x-A[h] | 1.88 | 0.50 | 0.22 | –[i] | 0 | 0.92 |
| | 7. 1xbi-A | 1sds-C | 0.33 | 0.62 | 0.56 | 0.40 | 1.14 | 0.79 |
| snRNA | 8. 1h64-1[h] | 1m8v-B[h] | 0.53 | 0.86 | 0.60 | 0.14 | –[i] | 0.90 |
| tRNA | 9. 1bs2-A | 1f7u-A | 3.44 | 0.20 | 0.17 | 1.10 | 1.33 | 0.85 |
| | 10. 1eqr-A | 1c0a-A | 1.64 | 0.30 | 0.69 | 1.06 | 1 | 0.75 |
| | 11. 1fmt-A | 2fmt-A | 1.17 | 0.43 | 0.80 | 1 | 1.67 | 0.87 |
| | 12. 1h3f-A | 1h3e-A | 9.48 | 0 | 0 | 0 | 0.12 | 0.60 |
| | 13. 1hc7-A[h] | 1h4s-A[h] | 1.28 | 0 | 0 | 0 | 0 | 0.80 |
| | 14. 1j09-A | 1n78-A | 1.87 | 0.42 | 0.45 | 0.59 | 0.64 | 0.91 |
| | 15. 1r5y-A | 1q2r-A | 0.72 | 0.38 | 0.43 | 0.93 | 1.27 | 0.67 |
| | 16. 1tui-A[h] | 1b23-P | 10.00 | 0.89 | 0.62 | 0.38 | 0.90 | 0.51 |
| | 17. 1u7d-A[h] | 1j1u-A[h] | 1.27 | 0.15 | 0.25 | 1 | 0 | 0.80 |
| vRNA | 18. 2ms2-A | 2bu1-A | 0.21 | 0.80 | 0.80 | 0.33 | 0 | 0.73 |

[a]See footnotes to Table 1, except that the gray background highlights predicted regions with $n_{TP}/n_{max} < 0.7$.
[b]PDB entry of the RNA-*free* protein structure.
[c]PDB entry of the RNA-*bound* protein structure.
[d]The root mean square deviation of the $C^{\alpha}$ atoms in the RNA-free protein structure relative to the respective RNA-bound protein structure.
[e]The ratio of $n_{TP}$ to $n^{max}$ in the patch or cleft predicted using the RNA-*free* protein structure.
[f]The ratio of $n_{TP}$ to $n^{max}$ in the patch or cleft predicted using the RNA-*bound* protein structure.
[g]The overlapping fraction, $f_{overlap}$, is computed according to Equation 5.
[h]The top-ranked patch is merged with nearby top-ranking clefts.
[i]The dash sign means that another disparate RNA-binding site could not be found.

tripeptide, –Gly–X–Gly–. As in previous studies (23), an aa with a relative SASA >5% is considered accessible for interacting with RNA, whereas that with a relative SASA ≤5% is deemed buried and inaccessible to a RNA molecule. The MOLMOL (24) program was used to compute the relative SASA of each aa from the protein structure using a solvent probe radius of 1.4 Å.

### Assignment of protonation states of ionizable residues

For a given RNA-binding protein, all Asp/Glu residues were deprotonated, while Arg/Lys residues were protonated. His residues were protonated if both side chain nitrogen atoms were within hydrogen bonding distance to any aa acceptor atom or water oxygen; otherwise they were assumed to be neutral, and the side chain nitrogen

that is within hydrogen bonding distance of an acceptor atom or water oxygen in the protein was protonated.

**Electrostatic ranking of each residue**

Each residue was assigned an 'electrostatic rank' (denoted as $Rank^{ele}{}_i$) based on whether it and its surrounding residues became electrostatically stabilized upon mutation to $Asp^-/Glu^-$. Thus, given the 3D structure of a *l*-residue RNA-binding protein, *l* mutant structures were generated by mutating each wild-type aa to $Asp^-/Glu^-$ depending on its size and shape. Ala, Asn, Asp, Cys, Gly, Ser, Thr, or Val were mutated to $Asp^-$, while the other residues were mutated to $Glu^-$. The side chain replacements were carried out using the SCWRL (25) program, which identifies the most common side-chain $\chi_1$ and $\chi_2$ angles for the mutant $Asp^-/Glu^-$ residue corresponding to the backbone $\phi$ and $\psi$ angles of the wild-type residue at that position. Each mutant structure was then energy minimized with heavy constraints on all non-hydrogen atoms using the AMBER (26) program to relieve bad contacts.

Having generated the *l* mutant structures, the gas-phase electrostatic energy of the wild-type ($E^{elec}{}_{wt}$) or mutant ($E^{elec}{}_{mut}$) protein in the *folded* state relative to that in an *extended reference* state ($E'^{elec}{}_{wt}$ or $E'^{elec}{}_{mut}$) was computed. In this extended reference state, the residues do not interact with one another, hence the electrostatic energy of the wild-type ($E'^{elec}{}_{wt}$) or mutant ($E'^{elec}{}_{mut}$) *unfolded* protein is simply the sum of the individual residue energies, and their difference is equal to the difference between the electrostatic energies of the native residue at position *i* ($E'^{elec}_i$) and the corresponding mutant $Asp^-/Glu^-$ residue ($E'^{elec}_{D/E}$). Thus, the change in the gas-phase electrostatic energy upon mutating aa *i* to $Asp^-/Glu^-$ is given by:

$$\begin{aligned}\Delta\Delta E^{elec}_i &= (E^{elec}{}_{mut,i} - E'^{elec}{}_{mut,i}) - (E^{elec}{}_{wt} - E'^{elec}{}_{wt}) \\ &= (E^{elec}{}_{mut,i} - E^{elec}{}_{wt}) + (E'^{elec}_i - E'^{elec}_{D/E})\end{aligned} \quad \mathbf{1}$$

A negative $\Delta\Delta E^{elec}{}_i$ implies that aa *i* is electrostatically stabilized upon mutation to an $Asp^-/Glu^-$. The gas-phase electrostatic energies were computed with the all-hydrogen-atom AMBER force field (27) with $\varepsilon = 1$ using the AMBER (26) program.

Knowing $\Delta\Delta E^{elec}{}_i$, the average electrostatic energy change of aa *i* and its surrounding, $<\Delta\Delta E^{elec}>_i$ was computed from:

$$< \Delta\Delta E^{elec} >_i = \sum \Delta\Delta E^{ele}{}_j / N^{aa}{}_i \quad \mathbf{2}$$

where the summation in Equation (2) is over $N^{aa}_i$ residues, which include aa *i* and all residues *j* whose $C^\alpha$ atoms are within 10 Å of the $C^\alpha$ atom of aa *i*. The *l* $<\Delta\Delta E^{elec}>_i$ values were then ordered from the most negative to the least negative/most positive and used to rank the *l* residues from 1 to 10 such that residues with the top 10% most negative $<\Delta\Delta E^{elec}>_i$ values were ranked 1, residues with the next 10% most negative $<\Delta\Delta E^{elec}>_i$ values were ranked 2, etc. (Supplementary Table S1).

**Evolutionary ranking of each residue**

Each residue was also assigned a 'conservation rank' (denoted as $Rank^{con}{}_i$) based on the relative conservation of the residue and its surrounding residues. For residue at position *i* in a given RNA-binding protein, a conservation score, $C_i$, was computed by the ConSurf program version 3.0 (19,28). The $C_i$ score reflects the evolutionary rate of the residue at position *i* in the phylogenetic tree generated on the basis of a protein's homologous sequences. The $C_i$ score is an integer number, ranging from 1 to 9, with 1 indicating a rapidly evolving and thus variable residue at position *i*, whereas 9, a slowly evolving, conserved residue.

Knowing the $C_i$ values, the average conservation of aa *i* and its surrounding, $<C>_i$, was computed from:

$$< C >_i = \Sigma C_j / N^{aa}{}_i \quad \mathbf{3}$$

where the summation in Equation (3) is over aa *i* and all residues *j* whose $C^\alpha$ atoms are within 10 Å of the $C^\alpha$ atom of aa *i*. Residues were then ranked from 1 to 10 such that residues with the top 10% largest $<C>_i$ values were ranked 1, residues with the next 10% largest $<C>_i$ values were ranked 2, etc. (Supplementary Table S1).

**Combined electrostatic and evolutionary ranking of each residue**

The $Rank^{ele}{}_i$ and $Rank^{con}{}_i$ values were multiplied to yield an overall ranking of each residue, denoted as $Rank_i$. Residues were ranked from 1 to 10 such that residues with the top 10% smallest $Rank^{ele}{}_i \times Rank^{con}{}_i$ values were ranked 1, residues with the next 10% smallest $Rank^{ele}{}_i \times Rank^{con}{}_i$ values were ranked 2, etc. (Supplementary Table S1).

**Generating surface patches**

Given the 3D structure of a *l*-residue RNA-binding protein, *l* irregular patches of various sizes were generated as follows: The $C_\alpha$ atom of each residue was chosen as the center of a patch and used to search for the nearest neighboring residue. If the latter residue has an overall electrostatic and evolutionary $Rank \leq 5$, it was included in the patch and its $C_\alpha$ atom defined a new center to search for the nearest neighboring residue. This process was repeated until the nearest neighboring residue had a $Rank > 5$. Patches containing at least 10 solvent accessible (surface) residues were considered as RNA-binding site candidates. However, if <3 patch candidates were found, the minimum number of surface residues in a patch was reduced by one successively until three or more candidates were found.

**Generating protein clefts**

Given the 3D protein structure, the 10 largest clefts (comprising cavities and grooves) were found using the SURFNET program (29). The SURFNET algorithm detects clefts; i.e. gap regions, by fitting spheres into spaces between any two atoms [see (29) for details]. If any atom of a residue was assigned as a constituent of the cleft by the SURFNET program, then this residue was regarded as a component of the cleft. When atoms of a residue were assigned to two different clefts, the residue was assigned to the larger of the two clefts. Residues constituting

a given cleft were removed if their overall *Rank* is >5. Clefts with 10 or more solvent accessible residues were considered as RNA-binding site candidates. However, if <3 cleft candidates were found, the minimum number of surface residues in the cleft was reduced by one successively until three or more candidates were found.

### Predicting RNA-binding residues

For each RNA-binding protein, two disparate RNA-binding regions were predicted, namely, a surface patch and a cleft. Each patch was scored using the $Rank^{ele}{}_i$ and $Rank^{con}{}_i$ values according to Equation (4):

$$< Rank^{ele}{}_i \times Rank^{con}{}_i > = \Sigma(Rank^{ele}{}_i \times Rank^{con}{}_i)/N^{patch}{}_i \quad \mathbf{4}$$

where the summation is over all $N^{patch}{}_i$ residues constituting patch *i*. The patch region with the smallest $<Rank^{ele}{}_i \times Rank^{con}{}_i>$ value (denoted by Patch[1]) is predicted to be the RNA-binding site. On the other hand, each cleft was scored according to the *mean* $C_i$ value of all the residues constituting the cleft. The cleft with the largest <*C*> value (denoted by Cleft[1]) is predicted to be another RNA-binding site.

To determine if the two predicted regions are disparate, the center of gravity of each region was determined, and the closest $C_\alpha$ atom was chosen as the respective center. If the distance between the Patch[1] and Cleft[1] centers is more than 10 Å, the two predicted RNA-binding regions are considered disparate. However, if this distance is ≤10 Å, Patch[1] and Cleft[1] were merged to yield a single RNA-binding region, and its gravity center was determined. In this case, another disparate RNA-binding candidate was determined by considering either the cleft with the next largest <C> or the patch with the next smallest $<Rank^{ele}{}_i \times Rank^{con}{}_i>$ value. The closest $C_\alpha$ atom to the gravity center of the RNA-binding candidate was chosen as the respective center, and its distance to the merged Patch[1] and Cleft[1] center was evaluated. If this distance is <10 Å, the procedure was repeated; otherwise it was halted. The putative RNA-binding residues are the solvent accessible residues in the two disparate RNA-binding regions; the total number of solvent accessible residues in each predicted RNA-binding region, $n_P$, is given in Table 1.

### Assessing the statistical significance of the predicted RNA-binding sites

For a given *l*-residue RNA-binding protein, if $n_{TP}$ of the $n_P$ surface residues are true positives, the statistical significance of such a prediction was assessed by computing the random pick value, (RPV), which is the probability of randomly picking a $n_P$-residue surface region whose number of true RNA-binding residues is greater than or equal to that in the predicted site. Thus, *l* 'random' patches with the same number of $n_P$ surface residues as the predicted patch or cleft were generated by choosing the $C_\alpha$ atom of each residue as an initial center to search for the nearest neighboring residue. The latter residue was included in the patch and its $C_\alpha$ atom defined a new center to search for the nearest neighboring residue. This process was repeated until the 'random' patch contained $n_P$ surface residues, which are assumed to bind RNA. Among the $n_P$ surface residues in each *random* patch, the number, $n_{TR}$, of true RNA-binding residues was counted and the maximum number, $n_{max}$, was recorded. The RPV value was then computed as the fraction of *random* patches with $n_{TR} \geq n_{TP}$. An RPV equal to zero means that zero chance of randomly picking a patch with $n_P$ surface residues containing greater than or equal to $n_{TP}$ true RNA-binding residues, whereas an RPV equal to one indicates no RNA-binding residues in the predicted RNA-binding region.

### Assessing the accuracy of the predicted RNA-binding sites

The strategy used to assess the accuracy of a given prediction is similar to that used in previous studies (18,30): the prediction for a given RNA-binding protein was deemed correct if the number of true positives, $n_{TP}$, in the predicted RNA-binding region is ≥0.7 $n_{max}$; i.e. $n_{TP}/n_{max} \geq 0.7$.

### Analyzing the overlap between the RNA-binding sites from the RNA-free and RNA-bound protein structures

The extent to which a RNA-binding site derived from the RNA-free protein structure overlapped with that derived from the corresponding RNA-bound structure was analyzed as follows. Let $n_P^{free}$ and $n_P^{bond}$ denote the total number of predicted RNA-binding residues derived from the RNA-free and RNA-bound protein structures, respectively, while $n_P^{overlap}$ is the number of predicted RNA-binding residues common to both binding sites. The overlapping fraction, $f_{overlap}$, is defined as

$$f_{overlap} = n_P^{overlap}/n_P^{free} \quad \text{if } n_P^{free} \leq n_P^{bound} \quad \mathbf{5a}$$

$$f_{overlap} = n_P^{overlap}/n_P^{bound} \quad \text{if } n_P^{bound} \leq n_P^{free} \quad \mathbf{5b}$$

For example, the total number of predicted RNA-binding residues based on the free 1a8v-A structure is 19, while that based on the RNA-bound 2a8v-B structure is 25. Since 13 predicted RNA-binding residues are common to both binding sites, the overlapping fraction, $f_{overlap} = 13/19 = 0.68$ (Supplementary Table S2).

## RESULTS

Given the 3D structure of a RNA-binding protein, our method predicted two disparate RNA-binding sites. To validate the method, it was first tested on 69 structurally non-homologous proteins whose structures have been solved in the presence of RNA. For each predicted site, Table 1 lists (i) $n_P$, the total number of *predicted* RNA-binding residues; (ii) $n_{TP}$, the number of predicted RNA-binding residues that truly bind RNA; i.e. the number of true positives; (iii) $n_{TP}/n_{max}$, the ratio of the number of true positives in the predicted site to the maximum number of true positives in randomly generated patches, each containing $n_P$ surface residues; and (iv) RPV, the probability of randomly picking a $n_P$ surface residue region whose number of true RNA-binding residues is greater than or equal to that in the predicted site. Figure 1 illustrates a predicted RNA-binding patch. For some small proteins, two disparate RNA-binding

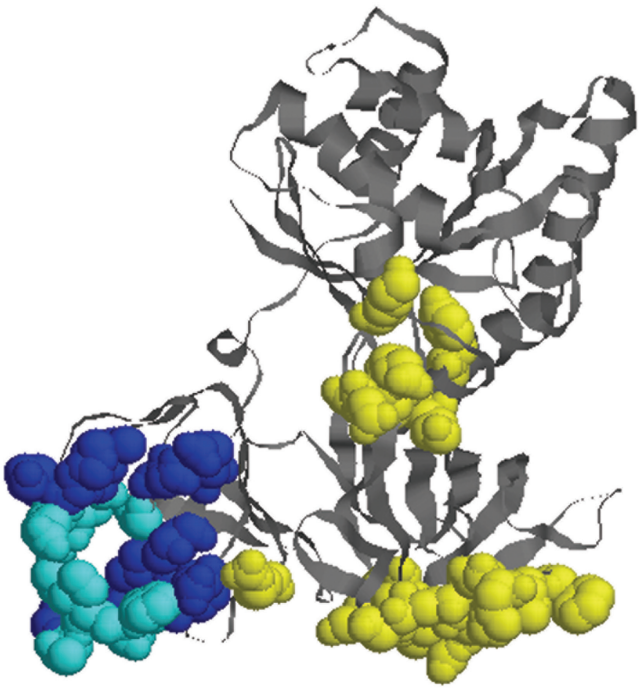

**Figure 1.** Predicted RNA-binding residues in the top merged patch + cleft (in blue) derived using the RNA-*free* elongation factor tu structure (1tui-A). The 'true' RNA-binding residues derived from the respective RNA-*bound* structure (1b23-P) structure are in yellow, while those that are correctly predicted are in cyan.

sites could not be found, and only a single RNA-binding site was predicted. Proteins with no correctly predicted region(s); i.e. $n_{TP}/n_{max} < 0.7$, are shaded.

### Predicting a single RNA-binding site

For a given RNA-binding protein, if we first choose the top-ranked surface patch and then a spatially disparate cleft, the first predicted RNA-binding region corresponds to either the top-ranked patch (if its center is more than 10 Å from the top-ranked cleft center) or the top-ranked patch merged with nearby top-ranking clefts (see Materials and Methods section). Among the 69 structurally non-homologous RNA-binding proteins, 65% (or 45/69) had a correctly predicted RNA-binding patch or patch + cleft with $n_{TP}/n_{max} \geq 0.7$; out of these, 47% (21/45) were statistically significant with RPV $\leq$0.1 (Table 1). A correctly predicted ribosomal protein (1vq8-1) exhibited a rather high RPV value of 0.79, indicating a 79% chance of randomly choosing a patch containing the same number of surface residues and $\geq$ number of true RNA-binding residues as the predicted patch. This is because this protein is essentially a RNA-binding domain with 91% (51/56) of the constituent residues involved in RNA binding. Three proteins, one bound to dsRNA (1yz9-A) and two bound to tRNA (1h3e-A, 1h4s-A) have predicted patches with RPV = 1, indicating no 'true' RNA-binding residues were predicted.

On the other hand, if we first choose the top-ranked cleft and subsequently a spatially disparate patch, the first predicted RNA-binding region corresponds to the top-ranked cleft alone or merged with nearby top-ranking patches (see Materials and Methods section). Based solely on the predicted cleft or cleft + patch for each protein, 59% (or 41/69) were correctly predicted with $n_{TP}/n_{max} \geq 0.7$ (Supplementary Table S3). Furthermore, 11 proteins possess no 'true' RNA-binding residues (RPV = 1) in the predicted cleft region; viz., 1yz9-A, 1fxl-A, 1wpu-A, 1feu-A, 1fjg-G, 1vq8-K, 1ec6-B, 1hq1-A, 1h4s-A, 1ddl-B and 2bu1-A. Thus, when only one RNA-binding site is predicted for each protein, choosing first a patch instead of a cleft yields better prediction accuracy.

### Predicting two disparate RNA-binding sites

Since RNA-binding sites are comprised of multiple disparate regions located on surface patches or clefts (see Introduction section), a cleft with highest mean conservation <C> that was spatially disparate from the top-ranked patch was also predicted as a RNA-binding site (see Materials and Methods section). Including a disparate RNA-binding cleft, in addition to the predicted patch, improved the accuracy of the RNA-binding residue prediction by 16%: out of the 69 test proteins in Table 1, 56 or 81% had at least one correctly predicted RNA-binding site with $n_{TP}/n_{max} \geq 0.7$. The RNA-binding site predictions for 11 RNA-binding proteins with insufficient true positives in the predicted patches were rescued by the cleft prediction. These correspond to the PDB chains 1m8x-A, 1fjg-K, 1fjg-T, 1g1x-B, 1mzp-A, 1sds-C, 1vq8-H, 1b23-P, 1c0a-A, 1f7u-A and 1q2r-A. For example, in the pumilio-homology domain complexed with mRNA, 1m8x-A, the predicted patch contains only two true positive RNA-binding residues, but the introduced cleft comprises 11 true positive RNA-binding residues with RPV = 0, meaning zero chance of randomly picking a patch with the same number of surface residues and $\geq$ number of true RNA-binding residues as the predicted cleft. However, two proteins (1yz9-A and 1h4s-A) did not contain any true RNA-binding residues (RPV = 1) in both predicted regions. Out of the 45 proteins with correct patch predictions, only 20 proteins also had correct predictions for the spatially disparate cleft. The above results indicate experimentally testing the top-ranked patch or merged patch + cleft before testing the spatially disparate cleft of a given protein for RNA binding.

For 44 proteins in Table 1, the top-ranked cleft and patch are spatially disparate, but for the remaining 25 proteins, the results may depend on the order in which the RNA-binding sites are chosen. The above accuracy of 81% was obtained by choosing first the top-ranked surface patch and then a spatially disparate cleft. If, instead, we chose first the top-ranked cleft and subsequently a spatially disparate patch, a comparable accuracy of 83% was obtained: 57/69 had at least one correctly predicted RNA-binding site with $n_{TP}/n_{max} \geq 0.7$ (Supplementary Table S3). Thus, when two disparate RNA-binding sites are predicted, the prediction accuracy is not too sensitive to the first choice of a patch or a cleft, in contrast to that for a single RNA-binding site.

### Effect of conformational changes upon RNA binding on RNA-binding site prediction

To evaluate how the predicted RNA-binding sites/residues would change when protein conformational changes upon RNA binding were neglected in using the RNA-free protein structures, the method was applied to a test set of 18 non-homologous proteins whose X-ray structures had been solved in the absence and presence of RNA. For the test proteins in Table 2, the $C_\alpha$ root mean square deviation (RMSD) of the RNA-free structure from the respective RNA-bound structure ranges from 0.21 to 10 Å. Regardless of the protein conformational changes upon RNA binding, the prediction accuracy based on the RNA-free and RNA-bound protein structures are comparable: 12 versus 10 of the 18 test proteins in Table 2 had at least one correctly predicted RNA-binding site with $n_{TP}/n_{max} \geq 0.7$ based on the RNA-free and RNA-bound protein structures, respectively.

To further assess the sensitivity of our method to protein conformational changes accompanying RNA binding, we identified those RNA-binding residues derived from the RNA-free and RNA-bound protein structures that are identical, and computed the overlapping fraction, $f_{overlap}$, according to Equation (5). The $f_{overlap}$ values in Table 2 indicate that more than half of the RNA-binding residues derived from the RNA-free and RNA-bound protein structures are identical for all but the 1ib2-A RNA-free protein. Notably, even when the $C_\alpha$ RMSD of the RNA-free protein, 1tui, from the respective tRNA-bound structure (1b23) is as large as 10 Å, half of the RNA-binding residues derived from the two structures are identical; also 13/18 proteins in Table 2 exhibit $f_{overlap} \geq 0.70$.

Analyses of the proteins with $f_{overlap}$ <0.7 show that the RNA-binding residues predicted from the bound states of three proteins are missing in the corresponding free structures. For example, residues 1141–1143, 1145–1150, 1152–1156, 1158–1159 and 1161–1168 comprise the patch predicted from the RNA-bound structure of pumilio-homology domain (1m8x-A), but residues 1150–1168 are missing in the respective RNA-free structure (1ib2-A); hence $f_{overlap} = 0.20$. Based on the RNA-bound structure of tyrosyl-tRNA synthetase (1h3e-A), the predicted RNA-binding cleft comprises of residues 79, 82, 83, 92, 144, 148, 149, 153, 167−171, 173, 174, 178, but residues 80–100 in the free structure (1h3f-A) are missing. Likewise, based on the RNA-bound structure of queuine tRNA-ribosyltransferase (1q2r-A), the predicted RNA-binding cleft is composed of residues 45, 47–49, 52, 73, 76–78, 92, 102, 106, 110, 111, 127, 128, 232–234, 260, 261, 264, 280–283, 286, 289, 290, 292 and 303, but residues 110 and 125–133 are missing in the free structure (1r5y). Interestingly, a stretch of missing residues in the free 1h3f-A (aa 85–100) and 1r5y (aa 125–133) structures are predicted to be disordered according to the neural network VLXT and/or VSL1 predictors of natural disordered regions (PONDR) (31,32). [Access to PONDR® was provided by Molecular Kinetics (6201 La Pas Trail-Ste 160, Indianapolis, IN 46268; 317-280-8737; E-mail: main@molecularkinetics.com). VL-XT is copyright© 1999 by the WSU Research Foundation, all rights reserved. PONDR® is copyright© 2004 by Molecular Kinetics, all rights reserved]. The above analysis shows that the binding sites predicted from the RNA-free and RNA-bound protein structures overlap if there are no disordered RNA-binding regions in the free structure that are ordered in the corresponding RNA-bound protein structure.

To further verify that our method is not too sensitive to protein conformational changes, the method was applied not only to the representative RNA-free structures, but also to the respective homologous structures of lower resolution. The results summarized in Supplementary Table S4 show that the prediction accuracy based on the representative structures is generally unchanged when lower-resolution homologous structures are employed, provided that residues missing in the representative X-ray structure are also missing at the respective positions in the homologous structures. In a few cases, however, the representative structure yielded no true RNA-binding residues ($n_{TP}/n_{max} = 0$), but the lower-resolution homologous structure yielded a correctly predicted site with $n_{TP}/n_{max} \geq 0.7$. One reason is due to missing RNA-binding residues in the representative structure that are resolved in the respective homologous structure. For example, RNA-binding residues 1156 and 1159 are missing in the 1.9 Å 1ib2-A representative structure, but are present in the 2.2 Å 1m8w homologous structure. Likewise, the side chains of RNA-binding residues 41 and 60 are missing in the 1wpv-A and 1wpt-A structures, respectively, but are present in the other homologous structures (1wrn, 1wro, 1vea, 1wps).

## DISCUSSION

In this article, we have developed a reliable method for predicting two disparate RNA-binding sites on a given RNA-binding protein based on detecting *evolutionarily conserved* residues located in (i) an electrostatically unstable surface patch and (ii) a cavity or groove. The method requires as input the protein structure and sufficient homologous sequences to define the relative conservation of each residue. It yields as output two sets of putative RNA-binding residues: the first set derived from a surface patch should be experimentally verified before testing the second set derived from a protein cleft. The method has the advantage of not being too sensitive to protein conformational changes upon RNA binding (Table 2 and Supplementary Table S4). On the other hand, it has the limitation of not being able to predict those RNA-binding regions whose folding is coupled with RNA binding, as these regions would be disordered in the free protein structure. This limitation, however, would be expected in all methods that depend on the protein structure in predicting RNA-binding residues. It may be alleviated using the PONDR (31,32) predictors to predict disordered segments in a protein from its sequence, but more RNA-free structures would be needed to test if this would indeed improve the prediction accuracy.

**Table 3.** Predicted RNA-binding residues in aminoacyl–tRNA–synthetases

| PDB-chain | Binding site | Predicted RNA-binding residues |
|---|---|---|
| 1h3f-A | Patch | I168–Y175, A178, Q179 |
| | Cleft[a] | L42, **G43–D45**, P46–D50, **H52**, **G54**, **H55**, **V58**, G77, F79, Y108, Q111, R155–D157, H171, E172, **Y175**, A178, **Q179**, **G194**, D196, **Q197**, N200, **P222**, **L223**, V225, R230, E231, **K232**, **S234**, **K235**, S236, I237, Y240, T244 –P246 |
| 1hc7-A | Patch[b] | S15, L19, Y30–T36, S88, E90, L91, **E113**, T114, **R142**, W143, **E144**, M145–R148, **L151**, **R152**, E155, **F156**, L157, **W158**, K199, K202–**F205**, A206, G207, **Q225**, **A226**, **T228**, **H230**, L232, N235, F236, **S258**, **G260**, **S262**, W263, **R264**, Q437, E438, T441, T443, A476, Y477 |
| | Cleft | I37–V39, Y44, L70, F71, F87, P89, A92, V93, V108, N139, V141, W143, E155, L157 |
| 1j09-A | Patch | L235, R237, N238, P239, D240, K241, T242, K243, I244, S245, K246, R247, K248, S249, H250 |
| | Cleft[c] | A7, **S9**, P10, T11, G12, D13, H15, G17, T18, I21, **E41**, D42, **T43**, **D44**, R45, A46, **R47**, V49, **K180**, Y184, T186, **Y187**, A206, E208, **W209**, L235, K243, I244, S245, K246, R247, S249, H250, S252, W255 |

[a]Residues in bold underlined are involved in binding ATP or tyrosinol based on the corresponding structure of tyrosyl-tRNA synthetase complexed with its cognate $tRNA^{Tyr}$, ATP and tyrosinol (1h3e-A).
[b]Residues in bold underlined are involved in binding ATP or prolinol based on the corresponding structure of prolyl-tRNA synthetase complexed with its cognate $tRNA^{Pro}$, ATP and prolinol (1h4q-A).
[c]Residues in bold underlined are involved in binding the tRNA 3′-terminal CytCytAde based on the structure of glutamyl-tRNA synthetase complexed with its cognate $tRNA^{Glu}$ and glutamol-AMP (1n78-A).

### Analysis of proteins with no apparent correctly predicted RNA-binding residues

One or more RNA-binding residues were correctly predicted using the RNA-free protein structures in Table 2, except for two proteins; viz., tyrosyl-tRNA-synthetase (TyrRS, 1h3f-A) and prolyl-tRNA-synthetase (ProRS, 1hc7-A). These two proteins are aminoacyl-tRNA-synthetases, which play a crucial role in protein synthesis by catalyzing the covalent coupling of its specific tRNA and aa in two steps. In the first step, the enzyme catalyzes the formation of an aminoacyl-adenylate (aa-AMP) from its cognate aa and ATP with the release of inorganic pyrophosphate. In the second step, the aminoacyl moiety is transferred from the aa-AMP to the 3′-terminal adenosine of the tRNA. Interestingly, for these two enzymes, the predicted sites correspond to the active-site pocket containing the aa-AMP intermediate.

For TyrRS (1h3f) and ProRS (1hc7), a predicted patch/cleft includes residues involved in binding ATP and tyrosinol or prolinol based on the structures of the corresponding tRNA-bound enzyme complexes (Table 3). Notably, in both the TyrRS/$tRNA^{Tyr}$ (1h3e) (33) and ProRS/$tRNA^{Pro}$ (1h4s) (34) complex structures, the tRNA 3′-end is disordered. However, the structure of the GluRS/$tRNA^{Glu}$/glutamol–AMP complex (1n78) (35) shows the tRNA 3′-terminal CytCytAde in the active site, within hydrogen bonding distance or vdW contact of residues 9, 41, 43, 44, 47, 107, 112, 116, 145, 177, 180, 181, 185, 187 and 209. Indeed, residues 9, 41, 43, 44, 47, 180, 187 and 209 were predicted to bind RNA based on the free glutamyl-tRNA-synthetase structure (1j09). Hence, some of the RNA-binding residues predicted using the free TyrRS (1h3f-A) and ProRS (1hc7-A) structures might be involved in binding the tRNA 3′-end.

### Comparison with other methods

As mentioned in the Introduction section, machine-learning approaches have also been used to predict RNA-binding sites (14–16). To compare the prediction accuracy of these methods with the present one, the methods to be compared have to be tested using the same set of proteins, the same definition of true RNA-binding residues, and the same accuracy criteria. Since the BindN server (14) is available, this method, which employs a support vector machine classifier, was compared with our method by using it to predict the RNA-binding residues in the 69 structurally non-homologous RNA-binding proteins. To reduce the number of false positive predictions, RNA-binding residues were predicted by setting the expected specificity to 80% in the BindN server. Since our method employs both sequence and structural information, whereas the BindN server uses sequence information only, it should yield more reliable predictions than BindN. Indeed, it predicts more true RNA-binding residues among the *predicted* ones than BindN: the ratio of true positives to *predicted* RNA-binding residues is 51% (1036/2026) using the present method, but 43% (1516/3496) using the BindN server (14).

## SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

## ACKNOWLEDGEMENTS

Funding for this work was provided by the National Science Council, Taiwan (NSC 95-2113-M-001-038-MY5). Funding to pay the Open Access publication charges for this article was provided by NSC 95-2113-M-001-038-MY5.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Tuschl,T. (2003) Functional genomics: RNA sets the standard. *Nature*, **421**, 268–272.
2. Draper,D.E. (1995) Protein-RNA recognition. *Annu. Rev. Biochem.*, **64**, 593–620.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
4. Cusack,S. (1999) RNA-protein complexes. *Curr. Opin. Struct. Biol.*, **9**, 66–73.

5. Draper,D.E. (1999) Themes in RNA-protein recognition. *J. Mol. Biol.*, **293**, 255–270.
6. Jones,S., Daley,D.T., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
7. Treger,M. and Westhof,E. (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recognit.*, **14**, 199–214.
8. Kim,H., Jeong,E., Lee,S.W. and Han,K. (2003) Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett.*, **552**, 231–239.
9. Morozova,N., Allers,J., Myers,J. and Shamoo,Y. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, **22**, 2746–2752.
10. Ellis,J.J., Broom,M. and Jones,S. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins*, **66**, 903–911.
11. Cai,Y.D. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, **1648**, 127–133.
12. Han,L.Y., Cai,C.Z., Lo,S.L., Chung,M.C. and Chen,Y.Z. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355–368.
13. Yu,X., Cao,J., Cai,Y., Shi,T. and Li,Y. (2006) Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.*, **240**, 175–184.
14. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
15. Jeong,E., Chung,I. and Miyano,S. (2004) A neural network method for identification of RNA-interacting residues in protein. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 105–116.
16. Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.
17. Kim,O.T.P., Yura,K. and Go,N. (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **34**, 6450–6460.
18. Chen,Y.C., Wu,C.Y. and Lim,C. (2007) Predicting DNA-binding sites on proteins from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. *Prot. Struct. Funct. Bioinform.*, **67**, 671–680.
19. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, 299–302.
20. Wu,C.Y., Chen,Y.C. and Lim,C. (2008) Specific RNA-binding structural motifs using a structural alphabet. In preparation.
21. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
22. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
23. Jones,S., Shanahan,H.P., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.*, **31**, 7189–7198.
24. Koradi,R., Billeter,M. and Wuthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 51–55.
25. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.J. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.
26. Case,D.A., Darden,T., Cheatham Iii,T.E., Simmerling,C., Wang,J., Duke,R.E., Luo,R., Merz,K.M., Pearlman,D.A. and Crowley,M. (2006) *AMBER 9*. University of California, San Francisco.
27. Duan,Y., Wu,C., Chowdhury,S., Lee,M.C., Xiong,G., Zhang,W., Yang,R., Cieplak,P., Luo,R., Lee,T. *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012.
28. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
29. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.
30. Jones,S. and Thornton,J. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.
31. Li,X., Romero,P., Rani,M., Dunker,A.K. and Obradovic,Z. (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform.*, **10**, 30–40.
32. Romero,P., Obradovic,Z., Li,X., Garner,E., Brown,C. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins.*, **42**, 38–48.
33. Yaremchuk,A., Kriklivyi,I., Tukalo,M. and Cusack,S. (2002) Class I tyrosyl-tRNA synthetase has a class II mode of cognate tRNA recognition. *EMBO J.*, **21**, 3829–3840.
34. Yaremchuk,A., Tukalo,M., Grotli,M. and Cusack,S. (2001) A succession of substrate induced conformational changes ensures the amino acid specificity of Thermus thermophilus prolyl-tRNA synthetase: comparison with histidyl-tRNA synthetase. *J. Mol. Biol.*, **309**, 989–1002.
35. Sekine,S., Nureki,O., Dubois,D.Y., Bernier,S., Chenevert,R., Lapointe,J., Vassylyev,D.G. and Yokoyama,S. (2003) ATP binding by glutamyl-tRNA synthetase is switched to the productive mode by tRNA binding. *EMBO J.*, **22**, 676–688.