Protein–RNA Interactions: Structural Analysis and Functional Classes

Jonathan J. Ellis,[1] Mark Broom,[2] and Susan Jones[1]*
[1]*Department of Biochemistry, School of Life Sciences, University of Sussex, Falmer, BN1 9RH, United Kingdom*
[2]*Department of Mathematics, School of Science and Technology, University of Sussex, Falmer, BN1 9RH, United Kingdom*

ABSTRACT A data set of 89 protein–RNA complexes has been extracted from the Protein Data Bank, and the nucleic acid recognition sites characterized through direct contacts, accessible surface area, and secondary structure motifs. The differences between RNA recognition sites that bind to RNAs in functional classes has also been analyzed. Analysis of the complete data set revealed that van der Waals interactions are more numerous than hydrogen bonds and the contacts made to the nucleic acid backbone occur more frequently than specific contacts to nucleotide bases. Of the base-specific contacts that were observed, contacts to guanine and adenine occurred most frequently. The most favored amino acid–nucleotide pairings observed were lysine–phosphate, tyrosine–uracil, arginine–phosphate, phenylalanine–adenine and tryptophan–guanine. The amino acid propensities showed that positively charged and polar residues were favored as expected, but also so were tryptophan and glycine. The propensities calculated for the functional classes showed trends similar to those observed for the complete data set. However, the analysis of hydrogen bond and van der Waal contacts showed that in general proteins complexed with messenger RNA, transfer RNA and viral RNA have more base specific contacts and less backbone contacts than expected, while proteins complexed with ribosomal RNA have less base-specific contacts than the expected. Hence, whilst the types of amino acids involved in the interfaces are similar, the distribution of specific contacts is dependent upon the functional class of the RNA bound. Proteins 2007;66:903–911. © 2006 Wiley-Liss, Inc.

Key words: protein–RNA interactions; RNA binding proteins; structural analysis; interaction propensity

INTRODUCTION

Proteins that interact with RNA molecules have diverse functions within the cell. Protein–RNA recognition is essential in the structure of the ribosome and spliceosome and plays important roles in gene expression. An understanding of how proteins recognize RNA is a key goal in structural biology, because it underpins many fundamental areas of molecular biology including gene splicing and viral replication. Currently, the mechanisms that control protein–RNA interactions are still poorly understood,

compared with those of protein–protein and protein–DNA interactions. One reason for this knowledge gap has been the small number of protein–RNA complexes for which structures have been solved. However, the publication of the structure of the 50S and 30S ribosome subunits in 2000,[1,2] and the advent of the structural genomics projects means that structural information for more than 350 protein–RNA complexes is currently available. This increased volume of data means that it is now possible to statistically analyze the structural and chemical characteristics of RNA binding sites, and make comparisons for proteins that bind different functional classes of RNA molecules, which is the focus of the current study.

Many studies have characterized protein–DNA interactions,[e.g.3–5] which have led to new methods for the prediction of DNA binding sites on protein structures.[e.g.6–8] As more structural data has become available, similar analyses have been conducted on protein–RNA complexes using increasingly large data sets of RNA binding proteins (RBPs) derived from the Protein Data Bank (PDB).[9–12] These analyses have shown that interactions that include nucleic acid backbone atoms and amino acid side chains predominate, and that arginine is a favored amino acid owing to the negative charge on the RNA. However, compared to DNA, RNA has greater structural diversity folding into A-form double helices or single strand motifs such as hairpins, loops and bulges.[13] This structural diversity has meant that previous analyses of protein–RNA complexes, conducted on data sets of more than one functional class of RNA, have shown variations in their characterization of RNA-binding sites.[9–12] Differences are observed for the favored amino-acid-base pairings, relative importance of hydrogen bonds and van der Waal (vdW), and specific hydrogen bond contacts.

The aim of the current study is to present a statistical analysis of the largest data set of protein–RNA complexes to date and make a novel comparison of binding site properties for proteins that bind to different functional classes of RNA. A greater understanding of how proteins interact with RNA will only be achieved if the structural and functional role of the RNA bound is considered. A knowledge of

*Correspondence to: Susan Jones, Department of Biochemistry, School of Life Sciences, University of Sussex, Falmer, BN1 9RH, United Kingdom. E-mail: s.jones@sussex.ac.uk

how the properties of interaction sites vary with the functional class of RNA bound will be essential if new methods are to be developed for RNA binding site prediction.

METHOD

A Protein–RNA Data Set

A preliminary data set of 348 protein–RNA complexes was extracted from the PDB[14]; the only criteria used to select these structures were that they contained both a protein and a RNA chain (as defined by the PDB search tool). These complexes were initially filtered to remove any structures solved by X-ray crystallography with a resolution < 3.0 Å and PDB entries where the protein or RNA chain contained less than five residues. The program HBPLUS[15] was used to calculate intermolecular hydrogen bonds and atom–atom contacts within the PDB files that remained. A maximum donor–acceptor distance of 3.35 Å and maximum hydrogen–acceptor distance of 2.7 Å were used to define a hydrogen bond. Atoms were considered to form vdW contacts if the distance between them was ≤3.9 Å and the contact had not been defined as a hydrogen bond. This set of PDB structures was further filtered, based on the results from HBPLUS, by removing any complex with <5 amino acid–nucleic acid contacts: this left 200 complexes in the data set. The results from HBPLUS also allowed a set of interacting pairs, one polypeptide and one polyribonucleotide chain, to be defined. The data set was then clustered into structurally homologous groups using the Structural Classification of Proteins (SCOP) database (version 1.67).[16] Chains with identical domains at the SCOP super-family level were clustered together, and the complex with the highest resolution was chosen to represent each cluster; NMR structures were only considered for representatives if the cluster contained no X-ray crystallographic structures. Structures that had no SCOP classification, of which there were 23, were removed from the data set.

The final data set consisted of 89 protein–RNA pairs (Table I). Each of these pairs were further clustered based on the molecular function of the RNA in the complex as defined by the Structural Classification of RNA (SCOR) database.[17] Of the RNA functional classes, only those containing at least five representative protein–RNA pairs were analyzed further. There were five such RNA functional groups: ribosomal RNA (rRNA), transfer RNA (tRNA), viral RNA (vRNA), messenger RNA (mRNA) and RNA as a protein ligand. This last group contains small RNAs, typically ≤10 nucleotides.

Defining Protein–RNA Interfaces

Two different approaches to define the protein–RNA interfaces were used: direct hydrogen bonding and accessible surface area (ASA). The program HBPLUS was used to calculate the number of direct hydrogen bonds between two chains (as described previously). This information was then used to define the interface as comprising all atoms that have a direct hydrogen bond to an atom in the RNA chain. The second method involved calculating the ASA of the interacting pairs and defining the interface

TABLE I. PDB Codes and Chain Identifiers of the Structurally Nonhomologous Protein–RNA Pairs Divided into Functional Classes Assigned by the SCOR database

RNA function	PDB codes
rRNA	1JJ21 1FJGM 1G1xB 1JJ2U 1MMSB 1MMSA 1G1xH 1N32G 1FJGT 1JJ2O
	1JJ2S 1JJ2A 1JJ22 1JJ2J 1FJGL 1JJ2B 1D6KA 1JJ2N 1JJ2I 1JJ2C
	1FJGB 1JJ2M 1JJ2X 1JBSA 1JJ2R 1I6UB 1JJ2E 1N32I 1FJGE 1FJGP
	1N32S 1JJ2W 1JJ2H 1FJGC 1JJ2Q 1G1xA 1FJGJ 1JJ2V 1FJGD 1JJ2D
	1MZPA 1JJ2T 1JJ2G 1FJGV
vRNA	1M8vI 1E6TA 1F6UA 1I9FB 1G70B 1MNBA 1A4TB
mRNA	1M8wB 1L1CA 1K1GA 1DZ5A 1CN8A 1KNZA
tRNA	1FFYA 1F7UA 1SERB 1IVSA 1N78A 1K8WA 1ASYA 1C0AA 1B23P 2FMTA
	1QTQA 1Q2RD 1J1UA 1H3EA 1QF6A 1H4SA
Ligand	2A8VB 1RC7A 2BBVC 1A34A 1GTFV 1D9DA 1AV6A 1R9FA 1UVJC

as those residues that lose ASA when the protein binds to the RNA. The ASA was calculated with the program Naccess (http://wolf.bms.umist.ac.uk/naccess). The ASA of the protein was calculated for the protein–RNA complex pair and for the protein in isolation. The difference in ASA between these two states was calculated for each residue, and any residue that lost ≥1 Å² was defined to be part of the RNA binding site of the protein.

Hydrogen Bonds and van der Waal Contacts

The contact data, derived from the HBPLUS program, were arranged in contingency tables where the rows were amino acids and the columns were nucleic acid moieties (each of the four bases, the ribose and the phosphate group). These tables allow χ^2 tests to be carried out with H_0 being that the observed distribution of bonds between the amino-nucleotide was random. The expected values for the hydrogen bond data were too small to perform a χ^2 test directly, so the data were collapsed in a number of ways including taking the totals of each nucleotide moiety without regard for the amino acid in the pairing. The hydrogen bond data were also combined with the vdW data resulting in a contingency table for all direct contacts. Although this type of test indicates whether the observed distribution of bonding is significantly different from the expected distribution, it does not provide any information about which of the pairings are different from expected. To calculate this, an analysis of the standardized residuals was carried out to indicate which of the parings were significantly different from the expected values ($\alpha = 0.05$). The standardized residual, d_{ij}, for the ijth

cell in an $i \times j$ table can be calculated by

$$d_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}} \cdot \sqrt{\frac{N - C_j}{N - Ri}},$$

where $E_{ij} = R_i C_j / N$, $N = \sum_{j=1}^{j} \sum_{i=1}^{i} n_{ij}$, $C_j = \sum_{i=1}^{i} n_{ij}$, and $R_i = \sum_{j=1}^{j} n_{ij}$. If N is large then $d_{ij} \approx \mathbb{N}(0, 1)$ and the significance of d_{ij} may be determined through standard tables.

Amino Acid Propensities

To analyze the relative importance of the different amino acids in the binding sites interface, propensity values were calculated by

$$P_{AA_j} = \frac{\sum_{i=1}^{Ni} \text{ASA}_{AA_{j(i)}} / \sum_{i=1}^{Ni} \text{ASA}_{(i)}}{\sum_{s=1}^{Ns} \text{ASA}_{AA_{j(s)}} / \sum_{s=1}^{Ns} \text{ASA}_{(s)}},$$

where $\text{ASA}_{AA_{j(i)}}$ is the contribution of amino acid j, in Å^2, to the RNA binding site; $\text{ASA}_{(i)}$ represents the total ASA of the RNA binding site (Å^2); $\text{ASA}_{AA_{j(s)}}$ is the contribution to the surface of the protein made by amino acid j excluding the binding site; and $\text{ASA}_{(s)}$ is the total ASA of the protein in isolation excluding the total ASA of the binding site, N_i is the number of residues in the binding site; and N_s is the number of residues on the protein surface excluding the binding site residues. The propensity function gives values centred around 1. A propensity value >1 indicates that a particular amino acid occurs more frequently in the RNA binding interface than on the surface of the protein. A propensity value <1 indicates that an amino acid occurs less frequently in the interface than on the surface of the protein. Propensity values were calculated for each amino acid using the complete data set of 89 protein–RNA pairs and for the five functional groups.

To ascertain whether a particular propensity value was significantly different from 1 (either above or below) a statistical bootstrapping method was implemented. This procedure involved simulating propensity values for each amino acid through random sampling of the ASA data sets used to create the actual propensity values. Any data set being affected by a small number of complexes will produce simulated propensities that vary greatly. Therefore, examining the distribution of these simulated propensities provides a measure of the variance contained in the data set. If the entire central 95% region (i.e., from the 2.5% point to the 97.5% point) is either above or below 1, there would be confidence ($\alpha = 0.05$) that the calculated value represents a true digression from 1. However, if 1 lies in the central 95% region, the calculated value cannot confidently be regarded as different from 1. For example, in the full data set, arginine has a propensity of 1.71, and the simulations provided a lower (2.5%) value of 1.55 with the upper (97.5%) value of 1.89. Therefore, 95% of the simulations are above 1 giving us confidence that the actual value calculated represents a true deviation from 1. In the case of phenylalanine, a propensity of 1.08 was calculated, and the upper and lower values were 0.79 and 1.44, respectively. These values sit either side of 1 and so indicate that we should not be confident that the value of 1.08 represents a propensity that is

TABLE II. Descriptive Statistics for Protein–RNA Interfaces in the Complete Data Set and for the Five Functional Classes

	Full data set	rRNA	tRNA	RNA ligand	vRNA	mRNA
Number	89	44	16	9	7	6
ASA						
Min.	120.8	443.3	183.9	182.1	293.6	539.9
Max.	7084.9	7084.9	2517.1	880.6	992.6	1084.3
Mean	1611.8					
SD	1284.8					
H-bonds						
Min.	1	6	2	1	3	8
Max.	139	139	44	20	12	30
Mean	25.5					
SD	23.8					

actually different from 1. This method, therefore, provides us with a way to determine whether the propensities calculated can be regarded as a significant deviation from 1.

Secondary Structure Motifs

At a higher level, the type of secondary structure present within the interface and the motifs they form are also important. Although some specific, but relatively large, folds that interact with RNA (e.g., the RNA recognition motif[18–20]) have been described, little is known about the secondary structure components of these interfaces. The program PROMOTIF[21] was used to calculate the secondary structure motifs (SSMs) of the interface residues. PROMOTIF defines 10 different SSMs; these are β-turns, γ-turns, helices (including both α- and 3_{10}-helices), β-strands, β-sheets, β-bulges, β-hairpins, βαβ units, ψ-loop and disulphide bridges. Propensity values for the SSMs were calculated in a similar manner to the amino acid propensities:

$$P_{\text{SSM}_j} = \frac{\sum_{i=1}^{Ni} \text{ASA}_{\text{SSM}_{j(i)}} / \sum_{i=1}^{Ni} \text{ASA}_{\text{SSM}_{(i)}}}{\sum_{s=1}^{Ns} \text{ASA}_{\text{SSM}_{j(s)}} / \sum_{s=1}^{Ns} \text{ASA}_{\text{SSM}_{(s)}}},$$

where $\text{ASA}_{\text{SSM}_{j(i)}}$ is the ASA that SSM j contributes to the interface; $\text{ASA}_{\text{SSM}_{(i)}}$ is the total ASA of the interface; $\text{ASA}_{\text{SSM}_{J(S)}}$ is the ASA that SSM j contributes to the surface excluding the interface; and $\text{ASA}_{\text{SSM}_{(s)}}$ is the total ASA of the surface excluding the interface. The SSM propensities give a measure of the relative importance of the different SSMs in the interface compared with the remainder of the protein surface. SSM propensities were calculated for the complete data set and for each of the RNA functional subcategories. The bootstrapping method (described previously) was also implemented on the SSM propensity data providing a measure of significance.

RESULTS

A data set of 89 nonhomologous protein–RNA pairs was extracted from the PDB and the size of the interfaces and the number of intermolecular hydrogen bonds involved in each complex are summarized in Table II. These data show a wide range of values for both characteristics for the whole data set and the five functional categories.

**TABLE III. Number of Observed Hydrogen Bonds Between Amino Acid and
Nucleotide Moieties for the Total Data Set**

Amino acid	Nucleotide moiety						Total
	A	C	G	U	Ribose	Phosphate	
Arg	13 (27.4)	34 (34.4)	51 (68.7)	24 (28.9)	155 (198.1)	366 (285.6)	*643*
Lys	11 (16.7)	17 (20.9)	29 (41.9)	11 (17.6)	81 (120.8)	243 (174.1)	*392*
Asn	7 (6.5)	9 (8.1)	8 (16.2)	22 (6.8)	58 (46.8)	48 (67.5)	*152*
Asp	0 (3.6)	6 (4.5)	27 (9.0)	4 (3.8)	38 (25.9)	9 (37.3)	*84*
Gln	9 (6.2)	9 (7.7)	22 (15.5)	13 (6.5)	54 (44.7)	38 (64.4)	*145*
Glu	8 (3.2)	4 (4.0)	17 (7.9)	2 (3.3)	38 (22.8)	5 (32.9)	*74*
His	6 (3.2)	3 (4.0)	4 (8.0)	4 (3.4)	36 (23.1)	22 (33.3)	*75*
Pro	2 (0.7)	1 (0.9)	2 (1.8)	0 (0.8)	12 (5.2)	0 (7.5)	*17*
Tyr	3 (2.6)	4 (3.3)	3 (6.5)	1 (2.7)	14 (18.8)	36 (27.1)	*61*
Trp	0 (0.7)	0 (0.9)	2 (1.8)	0 (0.8)	6 (5.2)	9 (7.5)	*17*
Ser	8 (7.2)	13 (9.0)	23 (18.1)	5 (7.6)	58 (52.1)	62 (75.1)	*169*
Thr	10 (5.7)	6 (7.1)	10 (14.2)	7 (6.0)	44 (41.0)	56 (59.1)	*133*
Gly	5 (5.0)	3 (6.3)	19 (12.6)	2 (5.3)	35 (36.3)	54 (52.4)	*118*
Ala	1 (2.0)	2 (2.5)	9 (4.9)	1 (2.1)	15 (14.2)	18 (20.4)	*46*
Met	3 (0.9)	1 (1.1)	5 (2.2)	0 (0.9)	8 (6.5)	4 (9.3)	*21*
Cys	2 (0.3)	0 (0.3)	0 (0.6)	2 (0.3)	0 (1.8)	2 (2.7)	*6*
Phe	0 (0.3)	0 (0.4)	2 (0.9)	0 (0.4)	4 (2.5)	2 (3.6)	*8*
Leu	4 (1.3)	4 (1.6)	0 (3.2)	1 (1.3)	15 (9.2)	6 (13.3)	*30*
Val	0 (0.9)	1 (1.1)	3 (2.2)	0 (0.9)	11 (6.5)	6 (9.3)	*21*
Ile	3 (0.6)	2 (0.8)	2 (1.6)	1 (0.7)	4 (4.6)	3 (6.7)	*15*
Total	*95*	*119*	*238*	*100*	*686*	*989*	*2227*

Numbers in parentheses are the expected values assuming a random distribution.

**TABLE IV. Number of Observed vdW Interactions Between Amino Acid and
Nucleotide Moieties for the Total Data Set**

Amino acid	Nucleotide moiety						Total
	A	C	G	U	Ribose	Phosphate	
Arg	463 (464.7)	402 (328.8)	397 (421.2)	322 (323.7)	1436 (1808.7)	1438 (1110.8)	*4458*
Lys	121 (243.9)	132 (172.6)	213 (221.1)	83 (169.9)	735 (949.4)	1056 (583.0)	*2340*
Asn	87 (108.7)	63 (76.9)	60 (98.6)	136 (75.7)	427 (423.2)	270 (259.9)	*1043*
Asp	32 (58.6)	43 (41.5)	81 (53.1)	18 (40.8)	327 (228.0)	61 (140.0)	*562*
Gln	39 (91.0)	61 (64.4)	91 (82.5)	74 (63.4)	422 (354.2)	186 (217.5)	*873*
Glu	65 (74.0)	61 (52.4)	73 (67.1)	69 (51.6)	361 (288.1)	81 (176.9)	*710*
His	196 (105.1)	38 (74.4)	58 (95.2)	99 (73.2)	493 (409.0)	124 (251.2)	*1008*
Pro	88 (71.8)	27 (50.8)	69 (65.1)	32 (50.0)	300 (279.5)	173 (171.7)	*689*
Tyr	120 (97.4)	70 (68.9)	74 (88.3)	172 (67.8)	332 (378.9)	166 (232.7)	*934*
Trp	67 (50.6)	54 (35.8)	116 (45.8)	36 (35.2)	165 (196.8)	47 (120.8)	*485*
Ser	77 (106.4)	77 (75.3)	108 (96.5)	45 (74.1)	426 (414.2)	288 (254.4)	*1021*
Thr	87 (84.3)	63 (59.7)	31 (76.4)	44 (58.7)	366 (328.2)	218 (201.6)	*809*
Gly	100 (124.5)	69 (88.1)	133 (112.8)	39 (86.7)	584 (484.4)	269 (297.5)	*1194*
Ala	59 (64.9)	42 (46.0)	48 (58.9)	39 (45.2)	285 (252.8)	150 (155.2)	*623*
Met	58 (39.5)	31 (28.0)	49 (35.8)	16 (27.5)	198 (153.8)	27 (94.4)	*379*
Cys	21 (6.2)	1 (4.4)	5 (5.6)	14 (4.3)	12 (23.9)	6 (14.7)	*59*
Phe	153 (72.8)	36 (51.5)	97 (66.0)	98 (50.7)	287 (283.2)	27 (173.9)	*698*
Leu	72 (53.9)	80 (38.1)	43 (48.8)	26 (37.5)	233 (209.8)	63 (128.8)	*517*
Val	60 (44.4)	25 (31.4)	28 (40.3)	10 (30.9)	227 (172.8)	76 (106.1)	*426*
Ile	32 (34.3)	38 (24.2)	36 (31.0)	19 (23.8)	156 (133.1)	47 (81.7)	*328*
Total	*1997*	*1413*	*1810*	*1391*	*7772*	*4773*	*19156*

Numbers in parentheses are the expected values assuming a random distribution.

Atom–Atom Contacts

The pattern of hydrogen bonds and vdW contacts between individual amino acids and nucleotides are summarized in Tables III and IV respectively. These data show that vdW interactions predominate. For each type of bond the sugar-phosphate backbone is preferred over the bases: 75% of hydrogen bonds involve the backbone, compared with 65% of vdW interactions. When the protein is considered the side chains show a stronger preference to be involved in RNA interactions than the main chain with
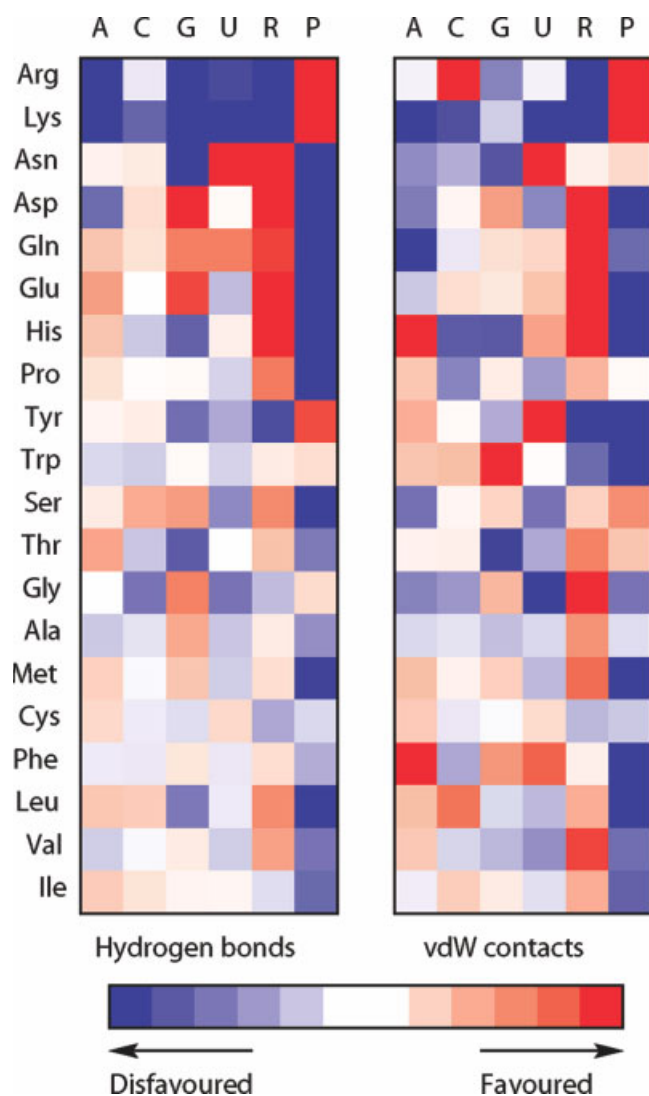
Fig. 1. These matrices show the pairing between each amino acid residue and nucleotide moiety. The colors indicate whether a particular pairing was observed more (red) or less (blue) than the expected. Higher color intensity indicates observations more extreme from the expected values than colors of lower intensity. The nucleotide moiety R indicates the ribose sugar and P the phosphate group.

27.9% of hydrogen bonds and only 25.9% of vdW contacts involving a protein main chain atom. When all four bases are grouped together, the distribution of hydrogen bonds between the nucleic acid moieties is 44% phosphate group, 31% ribose and 25% base (Table III). For vdW interactions, the distribution is 25% phosphate group, 41% ribose and 36% bases.

The expected values for hydrogen bond contacts in Table III are too small to perform a χ^2 on the table as whole; hence, the data were collapsed so only the nucleic acid moieties were considered. The χ^2 was significant ($\alpha = 0.05$) for this collapsed table indicating that the distribution of hydrogen bonds between the four bases and nucleic acid moieties is not random. Inspection of the totals shows that of the four bases guanine is the most preferred. However, all four bases have fewer hydrogen bonds than would be expected if the bonding was random. Hydrogen bonds to both the ribose and phosphate group are much more frequent with contacts to the phosphate group predominating. This indicates that there is a clear preference for hydrogen bonds to interact with the RNA backbone rather than the bases.

The hydrogen bond data in Table III and the vdW data in Table IV were combined to investigate the total bonding preference between amino acid residues and nucleic acid moieties (data not shown). A χ^2 test on the combined table showed that the bonding is not randomly distributed ($\alpha = 0.05$). An analysis of the residuals was also carried out, which lead to a list of amino acid-nucleotide moiety pairs that appeared more frequently and less frequently than the expected ($\alpha = 0.05$).

Figure 1 shows the pairings between each amino residue and nucleotide moiety for both hydrogen bonds and vdW contacts, and it shows that there is a similar pattern of pair preference for both types of contact. Two of the positively charged residues, arginine and lysine, have a clear propensity to bind to the phosphate group of nucleotides for both contact types with very small numbers of contacts between the other moieties. The other positively charged residue, histidine, does not follow the same pattern preferring to bind to the sugar group. Histidine appears to follow the pattern of aspartic acid, asparagine, glutamic acid,
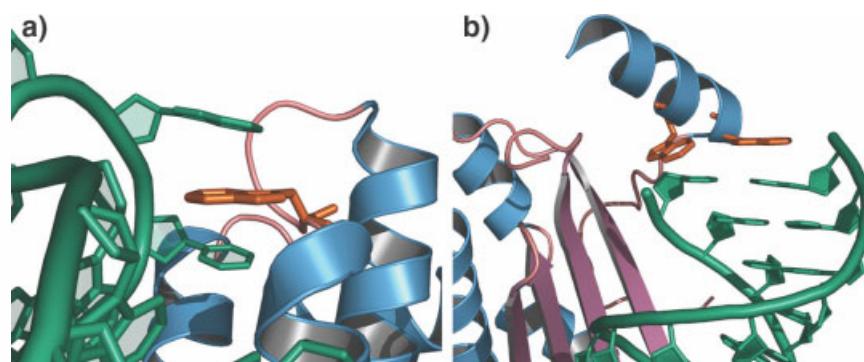

Fig. 2. Two examples of tryptophan stacking with RNA bases in protein–RNA complexes. (a) An arginyl-tRNA synthetase from *S. cerevisae* (1F7U[23]) Trp-569 (shown in orange) stacking in between I-934 (inosine) and C-935; (b) P19 complexed with 19-bp small interfering RNA from a tomato bushy stunt virus (1R9F[24]) Trp-42 and Trp-39 (shown in orange) stacking with C-1 of chain B and G-19 of chain C.

TABLE V. Number of Hydrogen Bonds Contacting Each Nucleotide Moiety Subdivided by RNA Functional Class

RNA group	Nucleotide moiety						Total
	A	C	G	U	Ribose	Phosphate	
rRNA	48 (63)	53 (84.3)	143 (160.5)	38 (63)	485 (491.2)	804 (708.9)	1571
tRNA	19 (15)	46 (20)	40 (38.1)	25 (15)	131 (116.6)	112 (168.3)	373
Ligand	2 (2.4)	6 (3.2)	10 (6.1)	2 (2.4)	18 (18.8)	22 (27.1)	60
vRNA	6 (2)	2 (2.7)	14 (5.2)	2 (2)	10 (15.9)	17 (23)	51
mRNA	11 (3.5)	8 (4.7)	12 (9)	19 (3.5)	26 (27.5)	12 (39.7)	88
Total	*86*	*115*	*219*	*86*	*670*	*967*	*2143*

Number in parentheses are the expected values Assuming a Random Distribution. The shaded values are observed frequencies that are 1.5 times greater or less than the expected frequencies.

TABLE VI. Number of vdW Interactions Contacting Each Nucleotide Moiety Subdivided by RNA Functional Class

RNA Group	Nucleotide moiety						Total
	A	C	G	U	Ribose	Phosphate	
rRNA	1013 (1282.4)	610 (914.6)	883 (1144.4)	544 (786)	5609 (5144.6)	3762 (3148.9)	12421
tRNA	363 (335.1)	433 (239)	353 (299.1)	204 (205.4)	1312 (1344.5)	581 (822.9)	3246
Ligand	21 (62.9)	137 (44.8)	93 (56.1)	27 (38.5)	255 (252.2)	76 (154.4)	609
vRNA	65 (64.3)	51 (45.9)	156 (57.4)	106 (39.4)	131 (258)	114 (157.9)	623
mRNA	424 (141.2)	114 (100.7)	198 (126)	275 (86.6)	259 (566.6)	98 (346.8)	1368
Total	*1886*	*1345*	*1683*	*1156*	*7566*	*4631*	*18267*

Number in parentheses are expected values assuming a random distribution. The shaded values are observed frequencies that are 1.5 times greater or less than the expected frequencies.

and glutamine, with the sugar group being preferred and the phosphate group clearly being disfavored. The nonpolar residues show less variation from their expected values—indicated by the lower color intensities in Figure 1. Notable exceptions to this are leucine and methionine, which both show a disliking for making contact with the phosphate group by either hydrogen bonds or vdW contacts. Phenylalanine also strongly disfavors making vdW contacts to the phosphate group.

Table V shows the observed and expected frequencies of hydrogen bonds to the different nucleic acid moieties for the five functional classes of RNA. Analysis of the observed frequencies that are 1.5 times larger or smaller than the expected frequencies reveals that those proteins binding to mRNAs, vRNAs, and tRNAs all have more contacts to two or more bases (cytosine, uracil and adenine), but less backbone contacts than expected. This is in contrast to proteins complexed with rRNAs, which show less contacts to two bases (cytosine and uracil) than expected. A similar trend is observed in the pattern of vdW contacts (Table VI), in which proteins contacting mRNAs and vRNAs show increased numbers of contacts to bases and reduced contacts to the backbone. Proteins contacting tRNA show increased binding to cytosine. In contrast, proteins contacting rRNA show reduced contacts to bases (specifically cytosine).

Amino Acid Propensities

The amino acid propensities for the full data set and for each of the RNA functional categories are shown in Table VII. The residues found to have a propensity signif-

icantly above 1 ($\alpha = 0.05$) are the positively charged residues arginine (1.71), lysine (1.18) and histidine (1.68), as well as serine (1.32), which is also polar. The two negatively charged amino acids are found to have the lowest propensities: glutamic acid (0.25) and aspartic acid (0.41). This was expected when the overall negative charge of the RNA is considered. However, two nonpolar residues, tryptophan (2.07) and glycine (1.37), were also found to have a propensity above 1. The amino acid propensities for the five functional categories of RNA follow a similar pattern to that of the whole data set.

SSM Propensities

Table VIII shows propensity values calculated for SSMs. The propensity values for helices and $\beta\alpha\beta$ units were found to be significantly below 1 in the full data set. From this data, $\beta\alpha\beta$ units were also significantly disfavored in the interface for the rRNA and tRNA functional groups. β-hairpins appear to be favored in the interface; although, only the values for tRNA and mRNA can be confidently regarded as different from 1.

DISCUSSION

In this study, 89 nonhomologous protein-RNA pairs have been analyzed, which include a total of 2227 hydrogen bonds and 19,156 vdW interactions. The mean size of an RNA binding site is 1611.8 Å^2 with a mean number of 0.016 hydrogen bonds per Å^2. It was observed that the binding sites were comprised of multiple disparate patches on the protein surface. A number of previous studies have analyzed protein-RNA interfaces using smaller

TABLE VII. RNAs Binding Amino Acid Propensities for the Full Data Set and the Five Functional Classes

Amino acid	Full data set	RNA functional category				
		rRNA	tRNA	Ligand	vRNA	mRNA
Arg	1.71	1.52	1.48	0.76	2.52	1.57
Lys	1.18	1.31	0.86	1.31	0.85	0.95
Asn	1.16	1.20	1.98	0.33	0.53	1.47
Asp	0.41	0.33	0.64	0.73	0.16	0.24
Gln	1.01	1.01	1.35	1.74	0.37	0.60
Glu	0.25	0.22	0.35	0.29	0.14	0.23
His	1.68	2.01	0.89	1.75	3.31	0.93
Pro	0.88	0.95	0.78	0.58	0.79	0.46
Tyr	1.26	1.11	1.54	1.74	1.27	2.76
Trp	2.07	2.26	2.26	4.77	0.64	18.78
Ser	1.32	1.62	0.99	1.61	1.17	1.53
Thr	1.04	1.14	1.21	0.97	0.44	0.97
Gly	1.37	1.35	1.12	1.54	0.82	1.14
Ala	0.84	0.86	0.85	0.54	1.17	0.10
Met	1.21	1.36	1.49	1.69	0.16	1.03
Cys	0.75	0.30	0.72	1.50	0.04	0.62
Phe	1.08	1.05	1.36	2.30	1.05	1.57
Leu	0.69	0.66	1.05	0.79	0.19	1.56
Val	0.80	0.81	0.98	0.98	0.18	0.84
Ile	0.85	0.73	1.47	0.34	0.55	0.72

The shaded values are propensities that were found to be significantly different from 1 at the 5% significance level.

TABLE VIII. RNAs Binding Secondary Structure Motif Propensities for the Full Data Set and the Five Functional Classes

Secondary structure	Full data set	RNA functional category				
		rRNA	tRNA	Ligand	vRNA	mRNA
β-turns	1.10	1.18	1.03	0.94	0.92	1.26
βαβ units	0.46	0.53	0.61	2.39	0.00	2.16
β-bulges	0.92	0.68	1.68	2.44	0.00	1.01
γ-turns	1.32	1.05	1.31	0.30	1.78	0.88
β-hairpins	1.23	1.08	1.43	2.34	0.42	2.16
Helices	0.85	0.86	0.97	0.72	1.48	0.92
ψ-loops	1.07	0.89	1.87	0.85	0.00	1.68
β-strands	1.00	0.91	1.20	2.52	0.46	1.61

The shaded values are propensities that were found to be significantly different from 1 at the 5% significance level.

data sets.[9–12,22] Their main conclusions and those of the the current work are summarized in Table IX. This table reveals that some trends in interface characteristics are consistent across all analyses, while others show disparities. All previous analysis conclude that (a) arginine and lysine are favored residues, (b) contacts to the RNA backbone are favored over contacts to base moieties and (c) proteins favor contacts through amino acid side chains. However, there are disparities between the relative importance of hydrogen bonds and vdW contacts, and the percentage of contacts to base, ribose and phosphate moieties. In addition, there are disparities between the preferences shown for specific bases and the most favored amino acid–base contacts.

The variations in the trends observed may reflect the increasing number of complexes analyzed and the inclusion of different proportions of proteins binding to different functional classes of RNA (a factor not addressed by any of the previous studies except Jones et al.[9]). In addition some disparities may reflect the variation in definitions of a protein-RNA contact. Some analyses only define RNA binding sites to include residues that make direct contacts through vdW and hydrogen bonds (and the definition of bond distances varies between studies), while others include all residues that contribute to the binding site by using loss of ASA to define the composition of the interface. In the current study, propensity calculations based on ASA showed that tryptophan is a favored residue in RNA binding sites, and the vdW data (Table IV) revealed that tryptophan has a clear preference for contacts to the double ringed nucleotides: guanine and to a lesser extent adenine. In a previous study the favored tryptophan–guanine pairing was

TABLE IX. Comparison of Protein–RNA Interface Characteristics as Presented in the Current Study and those of Jones et al.,[9] Treger and Westhof,[10] Jeong et al.[11] / Kim et al.,[22] and Lejeune et al.[12]

Parameter	Jones et al.	Treger et al.	Jeong et al./ Kim et al.	Lejeune et al.	Current study
Data set size	32 PDB entries	45 PDB entries	51 PDB entries	49 PDB entries	53 PDB entries (89 pairs)
Atom–atom contacts	vdW > H-bonds	vdW > H-bonds	—	H-bonds = vdW	vdW > H-bonds
NA-moiety H-bonds (B/R/P %)	—	—	49/28/23	35/43/22	25/31/44
Preferred base(s)	Guanine Uracil	No preference	Adenine Uracil	—	Guanine Adenine
AA-base	Arg-U Asn-G Asn-U Glu-G Gly-G	Arg-P Lys-P Met-P Phe-P Tyr-P	Arg-U Thr-A Lys-A Asn-U	Arg-C Arg-G Lys-C Arg-U Lys-A	Lys-P Tyr-U Arg-P Phe-A Trp-G
Favored AA	Lys, Tyr, Phe, Ile, Arg	Arg, Asn, Ser, Lys	Arg, Lys, Asn, Ser, Thr	Arg, Lys, Asn, His, Asp	Trp, Arg, His, Ser, Gly
Disfavored AA	Thr, Met, Leu, Gln, Trp,	Ala, Ile, Leu, Val	Trp, Phe, Gly, Met, Cys	Phe, Met, Trp, Glu, Cys	Val, Leu, Asp, Glu
Main/side Chain	—	Side > main	Side > main	—	Side > main
Backbone/base	BB > base	—	BB > base	BB > base	BB > bases

NA-moiety H-bonds are the percentages of each nucleotide moiety that were observed making hydrogen bonds to proteins. AA-base are the amino acid–nucleotide moiety pairs that are seen more often than the expected. Favored and disfavored AA are the residues that each study found to be favored or disfavored in protein–RNA interfaces. AA-base, favored amino acids and unfavored amino acids have all been limited to the five most/least favored. B: base; R: ribose; P: phosphate group; BB: backbone.

also observed.[10] It is possible that tryptophan has a role in base stacking with nucleic acids, and inspection of the complexes in the current study reveals some examples of this occurring (Fig. 2). Other studies of RBPs have also highlighted the importance of conserved tryptophan in rotaviruses,[25] although the exact function of these amino acids is not known. Another residue showing an unexpectedly high interface propensity was glycine. This has previously been observed by Treger and Westhof[10] whose analysis showed a preference for the glycine–guanine pairing. The presence of compact glycine residues could contribute to the flexibility of the interface a characteristic required to accommodate the complex and flexible RNA structures. The use of ASA for the calculation of interaction site residues emphasizes the importance of residues that do not make direct contacts to the RNA. In RNA binding, and indeed DNA and protein binding, it is the contributions made to the binding energy by all residues that define a stable complex. Hence, all interface residues need to be considered for accurate modeling of the interactions.

A large proportion of DNA binding sites on proteins are comprised of small compact motifs such as the helix-turn-helix[26] and the zinc fingers.[27] In contrast, RNA binding sites are more extensive and do not exhibit compact motifs for recognition. In the current analysis it was observed that the use of helices was disfavored in the recognition sites. This trend was observed qualitatively in the analysis by Jones et al.,[9] in which recognition through β sheets was recorded as the predominant mode of binding. Treger and Westhof[10] also found that RNA interface residues in helices that contacted RNA molecules through main-chain

contacts were less numerous than expected, and Draper[28] distinguishes two classes of RNA binding proteins, both of which include contacts through β structures. Nonhelical structures such as β-strands, β-hairpins and loops (observed as significantly favored in some functional categories in the current study) may occur more frequently due to their potential flexibility, which complements the flexible nature of the RNA structures bound.

The novel aspect of the current study is the comparison of interface parameters between proteins that form complexes with different functional classes of RNA. The amino acid propensities calculated for the functional classes showed trends similar to those observed for the complete data set (Table VII). However, the analysis of hydrogen bond (Table V) and vdW contacts (Table VI) revealed differences between the functional classes. In general, the proteins complexed with mRNA, tRNA and vRNA show a greater number of base specific contacts and fewer backbone contacts than expected, while the proteins complexed with rRNA show less base specific contacts than expected. The availability of bases for contacts with amino acid residues depends partly on the folding of the RNA. For example, the RNA in the ribosome subunits are comprised of predominantly dsRNAs; thus, the access of protein side chains to the base atoms is restricted.[29] An increased number of backbone contacts in the protein-rRNA complexes also fits with the current view that many of these proteins play a structural role in stabilizing the ribosome, and hence do not require base specific contacts. In contrast, the increased numbers of base contacts observed for proteins binding to mRNA, tRNA and vRNA reflects the fact that

these RNA molecules include large sections of ssRNA. For example, the two major sites of binding for aminoacyl-tRNA synthetases are the acceptor arm and the anticodon loop, which are both single-stranded substructures allowing sequence-specific contacts to be made. Similarly, many proteins complexed with mRNAs require sequence-specific contacts to function in such processes as polyadenylation, splicing and translation.

The current study shows that a clearer picture of protein-RNA interactions is beginning to emerge as the number of RBP structures increases. The comparison of interactions made to RNAs of different functional classes shows that the class of RNA bound is of importance in terms of the observed frequency of sequence specific and nonspecific contacts. It is intended that the trends observed in RNA binding sites will seed the development of tools to predict RNA binding function from protein structure data. In the light of the functional class data presented here, it is evident that more than one algorithmic approach will need to be developed for proteins bound to different classes of RNA.

REFERENCES

1. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 2000;289:905–920.
2. Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. Structure of the 30S ribosomal subunit. Nature 2000;407:327–339.
3. Luscombe N, Laskowski R, Thornton J. Amino acid-base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
4. Jones S, van Heyningen P, Berman H, Thornton J. Protein–DNA interactions: a structural analysis. J Mol Biol 1999;287:877–896.
5. Nadassy K, Wodak S, Janin J. Structural features of protein–nucleic acid recognition sites. Biochemistry 1999;38:1999–2017.
6. Tsuchiya Y, Kinoshita K, Nakamura H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins 2004;55:885–894.
7. Jones S, Barker J, Nobeli I, Thornton J. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res 2003;31:2811–2823.
8. Jones S, Shanahan H, Berman H, Thornton J. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 2003;31:7189–7198.
9. Jones S, Daley D, Luscombe N, Berman H, Thornton J. Protein–RNA interactions: a structural analysis. Nucleic Acids Res 2001;29:943–954.
10. Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recogn 2001;14:199–214.
11. Jeong E, Kim H, Lee S, Han K. Discovering the interaction propensities of amino acids and nucleotides from protein–RNA complexes. Mol Cells 2003;16:161–167.
12. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R. Protein–nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins 2005;61:258–271.
13. Cusack S. RNA–protein complexes. Curr Opin Struct Biol 1999;9:66–73.
14. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
15. McDonald I, Thornton J. Statifying hydrogen-bonding potential in proteins. J Mol Biol 1994;238:777–793.
16. Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
17. Klosterman PS, Tamura M, Holbrook SR, Brenner SE. SCOR: a structural classification of RNA database. Nucleic Acids Res 2002;30:392–394.
18. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. Science 1994;265:615–621.
19. Query CC, Bentley RC, Keene JD. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Cell 1989;57:89–101.
20. Oubridge C, Ito N, Evans PR, Teo CH, Nagai K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 1994;372:432–438.
21. Hutchinson E, Thornton J. PROMOTIF—a program to identify and analyze structural motifs in proteins. Protein Sci 1996;5:212–220.
22. Kim H, Jeong E, Lee S, Han K. Computational analysis of hydrogen bonds in protein–RNA complexes for interaction patterns. FEBS Lett 2003;552:231–239.
23. Delagoutte B, Moras D, Cavarelli J. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. EMBO J 2000;19:5599–5610.
24. Ye K, Malinina L, Patel DJ. Recognition of small interfering RNA by a viral suppressor of RNA silencing. Nature 2003;426:874–878.
25. Deo RC, Groft CM, Rajashankar KR, Burley SK. Recognition of the rotavirus mRNA 3′ consensus by an asymmetric NSP3 homodimer. Cell 2002;108:71–81.
26. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 2005;29:231–262.
27. Matthews JM, Sunde M. Zinc fingers–folds for many occasions. IUBMB Life 2002;54:351–355.
28. Draper D. Themes in RNA–protein recognition. J Mol Biol 1999;293:255–270.
29. Allers J, Shamoo Y. Structure-based analysis of protein–RNA interactions using the program ENTANGLE. J Mol Biol 2001;311:75–86.