Computational analysis of hydrogen bonds in protein–RNA complexes for interaction patterns

Hyunwoo Kim[a], Euna Jeong[a], Seong-Wook Lee[b], Kyungsook Han[a,*]

[a] *School of Computer Science and Engineering, Inha University, Inchon 402-751, South Korea*
[b] *Department of Molecular Biology, Dankook University, Seoul 140-714, South Korea*

Accepted 14 August 2003

First published online 27 August 2003

Edited by Robert B. Russell

Abstract Structural analysis of protein–RNA complexes is labor-intensive, yet provides insight into the interaction patterns between a protein and RNA. As the number of protein–RNA complex structures reported has increased substantially in the last few years, a systematic method is required for automatically identifying interaction patterns. This paper presents a computational analysis of the hydrogen bonds in the most representative set of protein–RNA complexes. The analysis revealed several interesting interaction patterns. (1) While residues in the β-sheets favored unpaired nucleotides, residues in the helices showed no preference and residues in turns favored paired nucleotides. (2) The backbone hydrogen bonds were more dominant than the base hydrogen bonds in the paired nucleotides, but the reverse was observed in the unpaired nucleotides. (3) The protein–RNA complexes contained more paired nucleotides than unpaired nucleotides, but the unpaired nucleotides were observed more frequently interacting with the proteins. And (4) Arg–U, Thr–A, Lys–A, and Asn–U were the most frequently observed pairs. The interaction patterns discovered from the analysis will provide us with useful information in predicting the structure of the RNA binding protein and the structure of the protein binding RNA.

Key words: Protein–RNA interaction; Hydrogen bond; Interaction propensity; Structural data

1. Introduction

RNA plays a key role in all the main functions of living molecules. It mediates the genetic information from the DNA to the protein, it adapts molecules for translation, and it is an integral enzymatic and functional component of the ribosome. These activities are all associated with the RNA binding proteins. Therefore, identifying how a protein binds to RNAs with specificity and affinity will provide insight into a wide range of biological processes.

A variety of problems concerned with protein–DNA complexes have been investigated for many years, but protein–RNA complexes have received much less attention in spite of their importance. One of the reasons for this is that there were a small number of protein–RNA complex structures available although an increasing number of protein–RNA structures have become known recently. In contrast to the regular helical structure of DNA, RNA molecules form complex secondary and tertiary structures consisting of stems, loops, and pseudoknots. The structural elements arranged into three-dimensional space are often recognized by specific proteins. RNA structures display hydrogen bonding, electrostatic, and hydrophobic groups that can interact with small molecules to form specific hydrogen bonds. However, it is unclear how the proteins interact with the RNA with specificity.

Analyzing the protein–RNA binding structures depends on a significant amount of manual work. Therefore, the protein–RNA binding structures are generally examined either individually or on a small scale. The task of analyzing the protein–RNA binding structures manually becomes increasingly difficult as the complexity and number of protein–RNA binding structures increase. In this study, we developed a set of algorithms for automatically analyzing the hydrogen bonds in protein–RNA complexes at the atomic level and for identifying the interaction patterns between the protein and RNA. A hydrogen bond is a relatively strong form of intermolecular attraction. A hydrogen bond plays a crucial role in the interactions of proteins and nucleic acid and in determining the tertiary and quaternary structures adopted by them.

In a previous study of 29 protein–RNA complexes [1], we extracted common features of protein–RNA complexes at the atomic level. As an extension of the previous study, we attempted to analyze the interaction patterns between the protein and RNA at the secondary structure level as well as the atomic level from a more comprehensive data set by means of statistical methods. The interaction patterns discovered from the analysis will provide us with useful information for predicting the structure of RNA binding protein and the structure of protein binding RNA.

2. Materials and methods

The overall framework for the structural analysis is given in Fig. 1. The structure data of the protein–RNA complexes were obtained from the Protein Data Bank (PDB) [2]. In order to remove homologous interactions in the data set, PSI-BLAST [3] was run on the protein–RNA complexes. From the remaining non-homologous, representative data set, the hydrogen bonds between the atoms of the amino acids and nucleotides were extracted. Secondary structure elements were assigned to each atom of the proteins and RNAs in the data set. The structure elements assigned to the proteins were helix, sheet, turn and others. Either 'paired' or 'unpaired' was assigned to the RNA nucleotides. The last process involved analyzing the hydrogen bonds and matching the secondary structures to the amino acids and nucleotides in the binding sites. A script generator was also de-

*Corresponding author. Fax: (82)-32-863 4386.
E-mail address: khan@inha.ac.kr (K. Han).

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00930-X

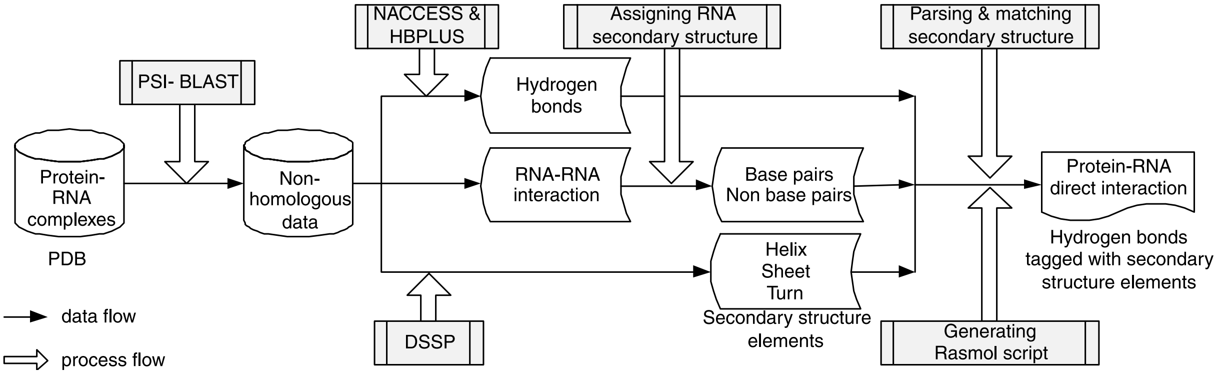

Fig. 1. Framework of the protein–RNA interaction analysis.

veloped for Rasmol (http://www.umass.edu/microbio/rasmol/) in order to make the binding patterns clear in terms of secondary structure elements.

2.1. Data set

The protein–RNA complexes, which were solved by X-ray crystallography with ≤3.0 Å resolution, were selected from PDB [2]. As of September 2002, there were 188 protein–RNA complexes in the PDB and the number of complexes with ≤3.0 Å resolution was 139. This study used PSI-BLAST [3] for the similarity search on each of the protein and RNA sequences in these 139 protein–RNA complexes in order to eliminate the equivalent amino acids or nucleotides in the homologous protein or RNA structures. After running the PSI-BLAST, the complexes were considered to contain representative and non-homologous interactions if they met the following conditions:

- The *E* value (the number of different alignments with scores equivalent to or better than the threshold, which are expected to occur in a database search by chance) is less than 0.001.
- The sequence identity (the extent to which two sequences are invariant) is below 80%.

Complexes whose proteins were homologous but recognized different nucleotide sequences were included in the data set. After running PSI-BLAST, 51 out of 139 protein–RNA complexes were left as the representatives. Table 1 shows the list of the 51 protein–RNA complexes.

2.2. Identification of hydrogen bonds

The number of hydrogen bonds between the amino acids and nucleotides in the protein–RNA complexes was calculated using HBPLUS [4]. The hydrogen bonds were identified by finding all the prospective atoms that satisfy given geometric criteria between the hydrogen bond donors (D) and acceptors (A). The positions of the hydrogen atoms (H) were theoretically inferred from the surrounding atoms, because the hydrogen atoms are invisible in purely X-ray-derived structures. The criteria considered to form the hydrogen bonds for this study are as follows: hydrogen bonds with a maximum D-A distance of 3.9 Å, a maximum H-A distance of 2.5 Å, and a minimum D-H-A angle as well as H-A-AA angle set to 90°, where AA is an acceptor antecedent.

The output file of HBPLUS includes information on the donor and acceptor atoms, computed distances and angles, etc. Meaningful protein–RNA bonds were extracted from the output files. This study analyzed the hydrogen bonds in terms of the interaction propensity, the atomic level property in the binding sites, the relationship between the main or side chain and the base or backbone hydrogen bonds, and the relationship between the secondary structures in the data set.

Table 1
The list of 51 protein–RNA complexes in the data set

PDB code							
1B23	1B2M	1B7F	1C0A	1C9S	1CX0	1DFU	1DI2
1DK1	1E7X	1EC6	1EFW	1F7U	1F8V	1FEU	1FFY
1FXL	1G59	1GAX	1GTF	1GTN	1G2E	1H4Q	1H4S
1HC8	1HDW	1HE0	1HE6	1HQ1	1I6U	1IL2	1JBR
1JBS	1JID	1K8W	1KNZ	1KQ2	1L9A	1LNG	1MMS
1QF6	1QTQ	1SER	1URN	1ZDH	1ZDI	2BBV	2FMT
5MSF	6MSF	7MSF					

2.3. Interaction propensity

The interaction propensity (*P*) was defined for each of the 20 most common amino acids binding each of the four nucleotides (adenine, guanine, cytosine, and uracil). The propensity function was based on that reported by Moodie et al. [5], but their propensity function was modified to determine the interaction propensity of the pairs between the amino acids and nucleotides on the surface. Amino acids on the surface were decided if the relative accessibility was larger than 5% according to the Naccess program (http://wolf.bms.umist.ac.uk/naccess). The interaction propensity P_{ab} between amino acid *a* and nucleotide *b* was defined by

$$P_{ab} = \frac{\sum N_{ab}/\sum N_{ij}}{(\sum N_a/\sum N_i)(\sum N_b/\sum N_j)} \tag{1}$$

where ΣN_{ab} is the number of amino acid *a* in hydrogen bonding to nucleotide *b*, ΣN_{ij} is the number of all amino acids in hydrogen bonding to any nucleotide, ΣN_a is the number of amino acid *a*, ΣN_i is the number of all amino acids, ΣN_b is the number of nucleotide *b*, ΣN_j is the number of all nucleotides, and all numbers were counted on the surface. The numerator $\Sigma N_{ab}/\Sigma N_{ij}$ represents the ratio of the co-occurrences of amino acid *a* and nucleotide *b* to the total number of all amino acids binding to any nucleotide on the surface. The first term $\Sigma N_a/\Sigma N_i$ of the denominator represents the ratio of the frequency of amino acid *a* to that of all amino acids on the surface, and the second term $\Sigma N_b/\Sigma N_j$ of the denominator represents the ratio of the frequency of nucleotide *b* to that of all nucleotides on the surface.

It should be noted that the interaction propensity of Eq. 1 was calculated as a proportion of a particular amino acid binding to a particular nucleotide on the surface divided by the proportion of them on the surface. Therefore, the propensity value can represent the frequency of the co-occurrences of amino acids and nucleotides in the protein–RNA complexes for each combination of amino acids and nucleotides. A propensity ≥1 indicates that a given amino acid occurs more frequently in the protein–RNA binding sites with a given nucleotide than on the remainder of the protein surface, whereas a propensity <1 indicates that a given amino acid occurs less frequently on the surface with a given nucleotide.

This propensity function is more refined than that used by Moodie et al. [5] and Jones et al. [6]. The interaction propensity value of Moodie et al. [5] is calculated as the proportion of a particular amino acid in the interface divided by the proportion of all amino acids in the interface. The residue interface propensity values used by Jones et al. [6] are similar to the interaction propensity value of Moodie et al. except that the accessible surface area of the protein calculated by Naccess was used instead of the number of residues. Both of their propensity values can calculate the frequency of each amino acid in the RNA binding sites, but they do not make a distinction between the nucleotides bound to an amino acid. Therefore, their propensity values cannot indicate what amino acid is favored in the binding sites with a particular nucleotide or interaction propensity of each nucleotide.

2.4. Secondary structure elements

The secondary structures of a protein were assigned using the DSSP program [7]. Given the atomic coordinates in PDB format, DSSP defined the secondary structure elements, geometrical features and solvent exposure of the proteins. The secondary structure elements defined by the DSSP were classified into four types: helix (α-helix, 3/10-helix, and π-helix), sheet (β = ladder and β-bridge), turn (hydrogen-bonded turn), and others (bend and other structures).

The RNA–RNA interactions were extracted from the HBPLUS output to assign a secondary structure element to each nucleotide. We considered two types of RNA secondary structure elements: paired and unpaired. If at least one hydrogen bond exists between the base atoms of two nucleotides, these two nucleotides were considered to be paired. If not, they were considered to be unpaired. If a hydrogen bond exists between other parts of two nucleotides than the base part, the two nucleotides were also considered to be unpaired. Watson–Crick base pairs, wobble base pairs, and all non-canonical pairs reported by Nagaswamy et al. [8] were classified as paired nucleotides.

3. Results

3.1. Frequency and propensity of direct hydrogen bonds

The hydrogen bonding interactions were calculated for the 51 protein–RNA complexes of Table 1. Table 2 shows the number of co-occurrences of amino acids and nucleotides and their interaction propensities. The amino acids in the first column are sorted by their average propensity values. In the data set, nucleotides A, G, C and U occurred 611, 956, 803 and 524 times, respectively. The total number of occurrences of the nucleotides was 2894. The propensity value 2.23 = (44/1204)/{(1600/20 618)(611/2894)} for the Arg–A pair, for example, was computed using Eq. 1.

Table 3 shows the number of direct hydrogen bonds between an amino acid and a nucleotide in the data set. The propensity value is not directly proportional to the number of hydrogen bonds. For instance, Trp has only 11 hydrogen bonds and its propensity value is 0.81, whereas Glu has 91 hydrogen bonds but its propensity value is as low as 0.78. The reason is that the propensity value depends on the total occurrences of residues on the surface of the protein.

As described earlier, the propensity value ≥ 1 indicates that a particular amino acid has a relatively high tendency to bind to a particular nucleotide on the surface. The average propensity values of amino acids are shown in the last column of Table 2. While amino acids revealed a diverse propensity values (ranging from 0.04 to 2.46), there is little difference in the propensity values of the nucleotides (ranging from 0.91 to 1.13). In general, hydrophilic residues have a high propensity value and hydrophobic ones have a low propensity. Guanine had the largest number of hydrogen bonds, but uracil revealed the highest propensity to interact with a protein.

The preferences for particular pairings between the amino acids and nucleotides are also shown in Fig. 2. For example, Arg–U, Asn–U, Lys–A, and Thr–A pairs occur frequently (Fig. 3). Since Arg has a long side chain with many electronegative atoms, it showed diverse binding patterns. Arg showed a high propensity binding to all four nucleotides. Among them, the Arg–U pair had the highest value and was discovered in the sxl-lethal protein (1B7F), hud and aurich RNA (1FXL, 1G2E), and *Escherichia coli* aspartyl-tRNA synthetase (1C0A, 1EFW).

Several amino acids strongly prefer specific nucleotides that can form a stable binding structure. Arg and Asn prefer uracil while Asp and Glu prefer guanine, and Thr particularly prefers adenine. Fig. 3 shows typical interaction patterns with two or more hydrogen bonds in the protein–RNA complexes. Fig. 3A,B shows the interaction patterns between Arg and uracil, while Fig. 3F,G shows the interaction patterns of Thr, which were found in most MS2–RNA virus complexes and some tRNA complexes. Fig. 3H is a stable and general pattern observed in many complexes such as ribonuclease, the trp RNA binding attenuation proteins, and some tRNA com-

Table 2
The frequency of the amino acids on the surface, frequency of the co-occurrences of an amino acid and a nucleotide, and the interaction propensities

Amino acid	Frequency (surface)	Frequency of co-occurrences					Propensity value				
		A	G	C	U	Total	A	G	C	U	Average
Arg	1 600	44	48	74	64	230	2.23	1.56	2.85	3.78	2.46
Lys	1 597	64	34	56	47	201	3.25	1.10	2.16	2.78	2.16
Asn	996	12	31	25	34	102	0.98	1.61	1.55	3.23	1.75
Thr	1 270	52	43	6	14	115	3.32	1.76	0.29	1.04	1.55
Ser	1 286	36	25	27	21	109	2.27	1.01	1.30	1.54	1.45
Tyr	615	11	8	16	12	47	1.45	0.67	1.61	1.85	1.31
Asp	1 322	2	52	33	8	95	0.12	2.04	1.54	0.57	1.23
Gln	782	4	18	11	16	49	0.41	1.19	0.87	1.94	1.07
Trp	232	2	0	6	3	11	0.70	0.00	1.60	1.22	0.81
Phe	614	0	29	0	0	29	0.00	2.45	0.00	0.00	0.81
Glu	1 996	10	55	19	7	91	0.41	1.43	0.59	0.33	0.78
His	552	4	9	2	8	23	0.59	0.85	0.22	1.37	0.71
Gly	1 648	6	16	13	3	38	0.30	0.50	0.49	0.17	0.39
Met	304	2	3	1	1	7	0.53	0.51	0.20	0.31	0.39
Cys	143	2	0	0	0	2	1.13	0.00	0.00	0.00	0.24
Pro	997	5	4	3	0	12	0.41	0.21	0.19	0.00	0.21
Leu	1 263	2	7	4	2	15	0.13	0.29	0.20	0.15	0.20
Ala	1 496	3	4	3	7	17	0.16	0.14	0.12	0.44	0.19
Ile	737	4	2	2	0	8	0.44	0.14	0.17	0.00	0.19
Val	1 168	0	0	3	0	3	0.00	0.00	0.16	0.00	0.04
Total	20 618	265	388	304	247	1204	1.04	0.98	0.91	1.13	

The frequencies of nucleotides A, G, C and U are 611, 956, 803 and 524, respectively. The total number of nucleotides is 2894. The propensity value 2.23 for the Arg–A pair, for example, was computed by (44/1204)/{(1600/20 618)(611/2894)}.

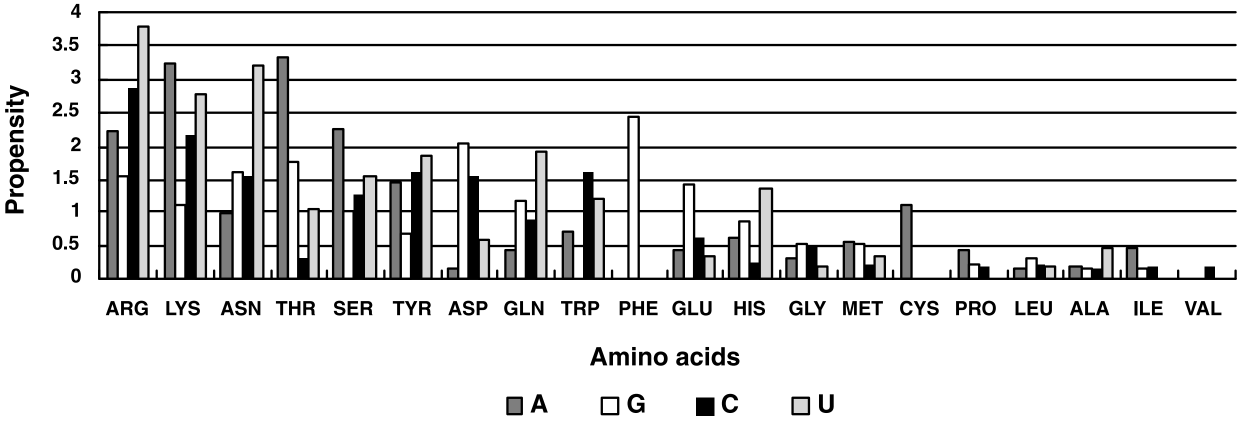

Fig. 2. Distribution of the propensity values of the direct hydrogen bonds between each of the 20 amino acids and each of the four nucleotides. The highest propensity value was observed for the Arg–U pair.

plexes. These interaction patterns form stable structures frequently observed, and have propensity values above average. For instance, the propensity of the Asp–G hydrogen bonds is 2.04, much higher than the average propensity 1.23 of Asp (Table 2). Glu, which also has an acidic side chain, has a similar tendency. Cheng et al. [9] recently generated 32 possible hydrogen bonding interactions between amino acid side chains and bases that involve two or more hydrogen bonds. Their interactions do not include the hydrogen bonding interactions in Fig. 3A–C while these patterns were observed frequently in our study.

3.2. Hydrogen bonding preferences between the main chain and side chain of the protein, and between the backbone and base of RNA

The atoms essential to hydrogen bonding exist both in the main chain and in the side chain of the proteins. Table 4 shows the observed frequency distribution of the hydrogen bonds in the main and side chain of each of the 20 amino acid residues. All the hydrogen bonds in the data set, including those on the surface, were computed at the atomic level. On average, side chain hydrogen bonds (71%) were observed more frequently than main chain hydrogen bonds (29%).

Table 3
The number of direct hydrogen bonds between amino acids and nucleotides in the data set

Amino acid	Nucleotide				Total
	A	G	C	U	
Arg	54	63	102	87	306
Lys	100	50	58	49	257
Ser	48	51	36	29	164
Thr	76	44	9	22	151
Glu	12	96	21	7	136
Asn	17	32	30	46	125
Asp	2	68	37	9	116
Gln	4	24	14	19	61
Tyr	13	10	21	15	59
Gly	6	18	13	3	40
His	8	15	4	9	36
Phe	0	31	0	0	31
Leu	3	7	7	2	19
Ala	3	4	3	7	17
Pro	5	4	3	0	12
Trp	3	0	6	3	12
Ile	6	2	2	0	10
Met	3	4	1	1	9
Cys	4	0	0	0	4
Val	0	0	3	0	3
Total	367	523	370	308	1568

If two hydrogen bonds exist between an amino acid and a nucleotide, the hydrogen bonds were counted twice.

No side chain hydrogen bonds were observed for the aliphatic residues (Ala, Gly, Ile, Leu, Phe, Pro, and Val) (Table 4). This is not surprising since they do not form side chain hydrogen bonds by their stereochemistry. More interestingly, these residues have 30% of the total main chain hydrogen bonds (132 out of 452) with a small number of hydrogen bonds (8% of the total 1568 hydrogen bonds). All of the aliphatic residues are also hydrophobic and have few electronegative atoms. In the hydrophobic residues, a shorter side chain resulted in a higher hydrogen bonding tendency. This is because the side chains of the hydrophobic residues may hinder the main chains from binding RNA. The hydrophobic residues except Trp and Tyr preferred main chain hydrogen binding to side chain hydrogen bonds. Other residues showed fewer main chain hydrogen bonds than side chain hydrogen bonds. Trp had a total number of 12 hydrogen bonds and all of these bonds were side chain hydrogen bonds. All aromatic residues (Trp, His and Tyr) with the exception of Phe preferred side chain hydrogen bonds to main chain hydrogen bonds.

Amino acids in which the side chain hydrogen bonds are more dominant than the main chain hydrogen bonds showed diverse interaction patterns. On the other hand, base and backbone hydrogen bonds were observed with almost equal frequency (49% for the base hydrogen bonds and 51% for the backbone hydrogen bonds) on average, as shown in Table 5. The two pyrimidines were observed to prefer backbone hydrogen bonds (64% for cytosine and 57% for uracil). The sugar hydrogen bonds of the four nucleotides had similar percentages.

3.3. Analysis of interactions at the secondary structure level

The protein–RNA complexes were analyzed to determine the interaction pattern in terms of secondary structures. The secondary structure elements of the RNA (paired nucleotides and unpaired nucleotides) were assigned by our own algorithm (algorithm not shown here), and the secondary structure elements of the protein (helix, sheet, turn and others) were assigned using DSSP [7].

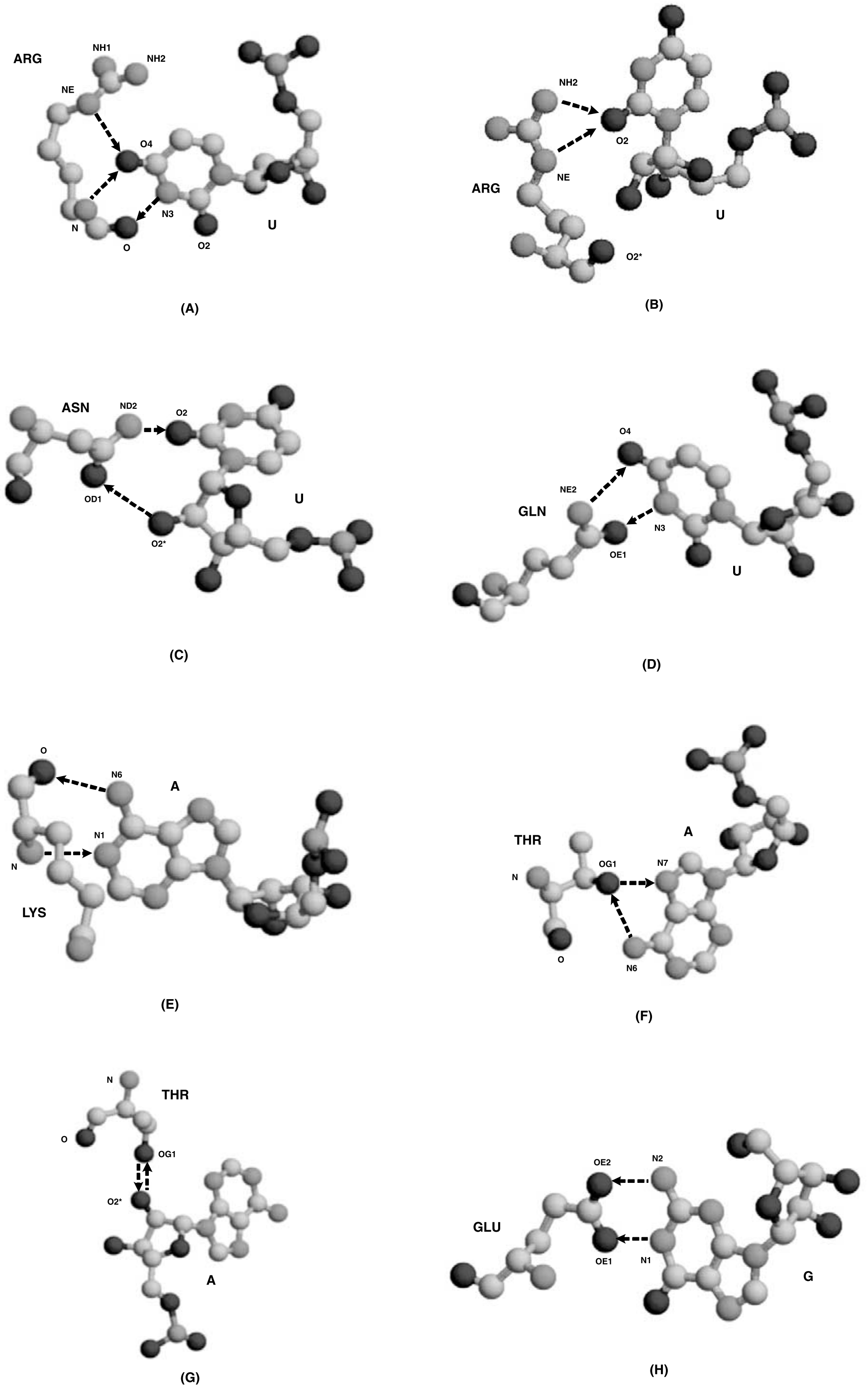

Fig. 3. Typical binding patterns between the amino acid and nucleotide pairs. The notation of the atoms follows the PDB format. The start of the arrows represents the donor and the end of the arrows represents the acceptor. All the interactions except (E) are hydrogen bonding interactions between amino acid side chains and nucleotides. The interaction (E) is the hydrogen bonding interaction between the main chain of Lys and adenine. The protein–RNA complexes that contains these interactions are as follows: (A) 1B7F, 1FXL, 1G2E, (B) 1QF6, 1IL2, 1B7F, (C) 1FEU, 1HC8, 1MMS (D) 1C0A, 1EFW, 1IL2, (E) 1C9S, 1GTF, 1GTN (F) 1E7X, 1HDW, 1HE0, 1HE6, 1ZDH, 1ZDI, 5MSF, 6MSF, 7MSF (G) 1B23, 1GAX, 1I6U, (H) 1B2M, 1H4Q, 1C9S, 1GTF, 1GTN.

In order to understand the relationship between the hydrogen bonding locations to the amino acids (i.e. base or backbone) and the corresponding secondary structure elements (i.e. paired or unpaired), the ratio of the number of observed base and backbone hydrogen bonds was calculated. While the backbone hydrogen bonds were more dominant than the base hydrogen bonds in the paired nucleotides, these preferences were reversed in the unpaired nucleotides, as shown in Table 6.

The total number of nucleotides in the data set is 2894. The number of paired nucleotides is 2001 (70%) and the number of unpaired nucleotides is 893 (30%). Therefore, the ratio of paired nucleotides to unpaired nucleotides was 7 to 3. However, the ratio became 4 to 6 in the RNAs interacting with the

Table 4
Analysis of the hydrogen bonds in the main or side chain of the amino acids

Amino acid	Main chain		Side chain		Total
	Main (M)	M%	Side (S)	S%	M+S
Ala	17	100	0	0	17
Arg	32	10.46	274	89.54	306
Asn	22	17.60	103	82.40	125
Asp	26	22.41	90	77.59	116
Cys	2	50	2	50	4
Gln	2	3.28	59	96.72	61
Glu	10	7.35	126	92.65	136
Gly	40	100	0	0	40
His	3	8.33	33	91.67	36
Ile	10	100	0	0	10
Leu	19	100	0	0	19
Lys	92	35.80	165	64.20	257
Met	6	66.67	3	33.33	9
Phe	31	100	0	0	31
Pro	12	100	0	0	12
Ser	37	22.56	127	77.44	164
Thr	72	47.68	79	52.32	151
Trp	0	0	12	100	12
Tyr	16	27.12	43	72.88	59
Val	3	100	0	0	3
Total	452	28.83	1116	71.17	1568

M and S stand for the number of hydrogen bonds in the main chain and side chain of each amino acid respectively. M% and S% represent the percentages of hydrogen bonds and were computed by M/(M+S) for the main chain and S/(M+S) for side chain, respectively.

proteins. It follows from this observation that the unpaired nucleotides are more flexible than paired nucleotides to interact with the proteins.

The observed distribution of the secondary structure elements in the binding sites is shown in Table 7. This table shows the number of hydrogen bonds observed in each combination of protein–RNA bindings at the secondary structure level and the ratio of the paired nucleotide hydrogen bonds to the unpaired nucleotide hydrogen bonds in each protein in the secondary structure. It is interesting that the sheet residues have a much higher tendency to recognize unpaired nucleotides than the residues in the helix, turn, and others, as shown in Table 7.

Adenine and uracil are similar to each other in the ratio of paired nucleotides to unpaired nucleotides, but they are very different when hydrogen bonding to amino acids (Table 7). The P/NP ratio for uracil is 0.82, which is the highest value among the four nucleotides, while that for adenine is 0.24, which is the lowest. This means that paired uracil frequently hydrogen bonds to amino acids while paired adenine hardly bonds to amino acids.

Pictorial representations of the two canonical hydrogen bonds are shown in Fig. 4. Fig. 4A shows the hydrogen bonds between the helix residues and the base-paired nucleotides in seryl-tRNA synthetase (1SER). Since the bases form pairs, the bases are towards the inside of the RNA and the backbones are toward the opposite. Backbones are relatively free to hydrogen bonding to proteins. In this conformation, the α-helix residues are easily inserted into the RNA helices. This binding pattern is often shown in DNA–protein recognition [10,11]. The other pattern was observed in the sheet residues binding the unpaired nucleotides. The interaction pattern in Fig. 4B was observed in the MS2 coat protein complex (7MSF). The amino acid side chains are exposed parallel to the direction of the sheet residues. Since the sheet residues have a high tendency to bind to the unpaired nucleotides, the protein binding site is not limited to any particular nucleotide base and backbone. These two binding patterns corroborate the results of previous reports on protein–RNA recognition [11].

The RNA secondary structure elements in the binding sites are summarized in Table 8. Arg, which interacts most frequently with the RNAs, has a moderate and balanced binding frequency to each of the four nucleotides. Lys shows a strong tendency to bind to the unpaired nucleotides when compared to the paired nucleotides, particularly with adenine. In addition to Lys–unpaired adenine, Glu–unpaired guanine were also observed frequently.

3.4. Related work

This section first compares the results of our analysis with those reported by Jones et al. [6] (referred to as Jones from now on), and with other studies. The results were compared in four aspects: (1) the contact preference between the bases and backbones in the RNA, (2) which bases are better recognized by a protein, (3) which residues have a high propensity to bind to RNA, and (4) the contact preferences between a particular amino acid and a particular nucleotide.

The ratio of the base to backbone contacts (0.96) in this study is similar to that (1.00) in Jones' study. In our analysis, uracil was the most preferred nucleotide, which was followed by adenine. However, in Jones' work, guanine and uracil were the most preferred. The favored residues in our analysis are Arg, Lys, Asn, and Thr in decreasing order of interaction propensity. In Jones' analysis, Arg and Tyr were the most favored. Although the ranking of the favored residues between two studies was different, Arg and Tyr showed a high propensity to bind to RNA in both studies. The high interaction preferences between the particular pairs were observed in Arg–U, Thr–A, Lys–A, Asn–U, Arg–C, and Lys–U in our analysis. In Jones' analysis, Arg–U, Arg–phosphate, Asn–G, Asn–U, Glu–G, Gly–G, Thr–A, and Tyr–sugar were the pre-

Table 5
Analysis of the hydrogen bonds in the base and backbone of the nucleotides

Nucleotide	Base		Backbone					Total
	B	B%	P	P%	S	S%	P%+S%	B+P+S
A	196	53.41	86	23.43	85	23.16	46.59	367
G	305	58.32	65	12.43	153	29.25	41.68	523
C	134	36.22	123	33.24	113	30.54	63.78	370
U	131	42.53	90	29.22	87	28.25	57.47	308
Total	766	48.85	364	23.21	438	27.93	51.15	1568

B, P, and S stand for the number of hydrogen bonds in the base, phosphate, and sugar respectively. B%, P%, and S% for the percentages of hydrogen bonds in each part.

Table 6
The numbers of base and backbone hydrogen bonds, and the ratio of base to backbone hydrogen bonds for each of the four nucleotides at the RNA secondary structure level

H-bond	Nucleotide										
	A		G		C		U		Total		
	P	NP	P	NP	P	NP	P	NP	P	NP	P+NP
Base	12	184	92	213	17	117	27	104	148	618	766
Backbone	58	113	132	86	145	91	112	65	447	355	802
Base+backbone	70	297	224	299	162	208	139	169	595	973	1568
Base/backbone	0.21	1.63	0.70	2.48	0.12	1.29	0.24	1.60	0.33	1.74	0.96

P: paired nucleotide, NP: unpaired nucleotide, base/backbone: ratio of the base hydrogen bonds to backbone hydrogen bonds.

ferred pairs. However, Gly–G showed a very low interaction propensity in this study. The preferences between the amino acid residues and phosphate or sugar could not be calculated in this study since the interaction propensity was designed for each combination of the 20 amino acids and four nucleotides.

In summary, the analysis in this study showed similar results in several respects, but did not precisely correspond to their results. This can be explained by two reasons: (1) difference in the interaction propensity functions, and (2) difference in the data sets. While their propensity function measures the tendency of each amino acid residue to occur in the RNA binding site, our propensity function measures the binding tendency of each amino acid to each of the four nucleotides via a hydrogen bond. Therefore, in Jones' study a residue in the interface can have a high propensity value even if it does not actually bind to RNA. However, a residue in the interface does not contribute to our propensity value unless it is involved in hydrogen bonding with RNA. While the propensity function of Jones' study considers both hydrophobic interactions and hydrogen bonds, our propensity function considers hydrogen bonds only.

Our interaction propensity function has a few advantages. First, it can tell the binding propensity of an amino acid with *each* of the nucleotides, but their propensity function cannot do this simply because it does not distinguish nucleotides binding with an amino acid. For example, Glu often forms stable binding to guanine, as shown in Fig. 3H, but this type of binding pattern cannot be discovered by the propensity function of Jones' study since it does not distinguish the nucleotides binding to a residue. Second, hydrophobic interactions are much weaker than hydrogen bonds but they are considered with equal importance in the propensity function of Jones' study. Therefore, the interaction propensity values of hydrophobic amino acids are computed higher in Jones' study than expected.

Regarding the difference caused by the use of different data sets, 51 non-homologous protein–RNA complexes were analyzed in this study, whereas Jones' study examined only 20 complexes. Basically, our data set includes those reported by Jones in terms of the family level and other complexes. To explain the difference caused by different data sets, we computed our interaction propensity function on Jones' data sets of 20 complexes, and Fig. 5 summarizes the result. While the propensity values of this data set ranged from 0.2 to 1.7 by Jones' propensity function, the values ranged from 0.0 to 2.63 in ours, which has more spread distribution. In Fig. 5, the propensity values of hydrophobic residues such as Ile and Val are as low as almost 0, whereas they were reported higher than 1.0 in Jones' study. This is because the propensity function of their study computes a high value for hydrophobic residues, which have a high tendency to exist in the interface rather than being involved in hydrogen bonds. On the other hand, hydrophilic residues such as Arg and Lys have many electronegative atoms and therefore are frequently involved in hydrogen bonds. Such hydrophilic residues have higher propensity values by our function than by Jones' function. Asp is a hydrophilic residue but has a very low value in Jones' study, whereas it has a value above average by our function, which seems a more reasonable value.

Although the results of our analysis show differences from Jones' study, our results agree with many previous studies. In the study by Nobeli et al. [12], Tyr and Thr prefer adenine to guanine but Glu and Asp prefer guanine to adenine. This discrimination between adenine and guanine is also shown in our study. In Table 2, the propensity values of Thr–A and Thr–G are 3.32 and 1.76, respectively, whereas those of Asp–G and Asp–A are 2.04 and 0.12, respectively. The reason that Asp and Glu favor guanine is that they form stable binding structures with guanine, as shown in Fig. 3H.

In the study by Treger and Westhof [13], the ratio of the base to backbone bonds is 0.31 in helices but the ratio is 0.64 in sheets. This result is in agreement with our observations

Table 7
Distribution of hydrogen bonds observed in each combination of protein–RNA bindings at the secondary structure level

H-bond	Nucleotide														
	A			G			C			U			Total		
	P	NP	P/NP	P	NP	P/NP	P	NP	P/NP	P	NP	P/NP	P	NP	P/NP
Helix	24	53	0.45	70	36	1.94	65	83	0.78	35	14	2.50	194	186	1.04
Sheet	21	175	0.12	36	135	0.27	28	57	0.49	27	94	0.29	112	461	0.24
Turn	5	21	0.24	50	12	4.17	27	16	1.69	22	11	2.00	104	60	1.73
Others	20	48	0.42	68	116	0.59	42	52	0.81	55	50	1.10	185	266	0.70
Total	70	297	0.24	224	299	0.75	162	208	0.78	139	169	0.82	595	973	0.61

P: paired nucleotide, NP: unpaired nucleotide, P/NP: ratio of P to NP.

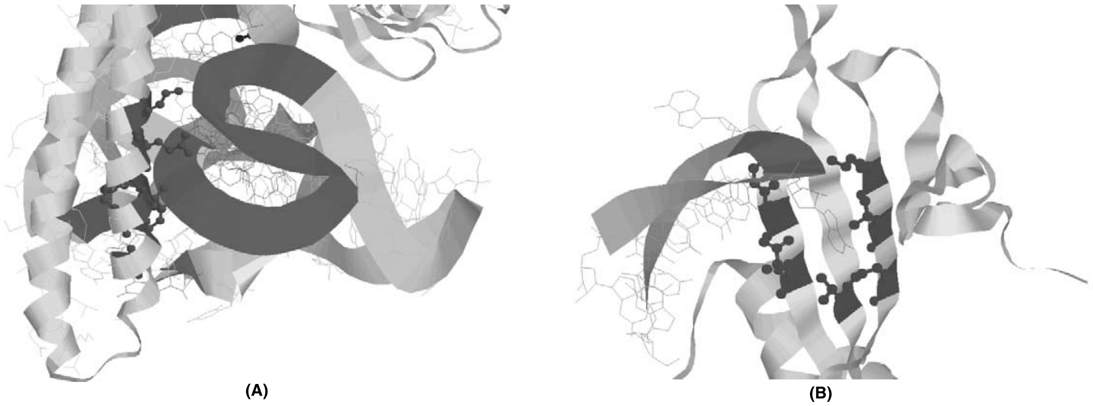

Fig. 4. Special binding patterns between (A) the helix residues and nucleotides in the base pairs, and (B) sheet residues and the unpaired nucleotides. Color: green ribbon for the nucleotides, red ribbon and stick for the α-helix residues, blue ribbon and stick for the sheet residues, and the dark color for the binding part. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that sheets prefer unpaired nucleotides and that the base hydrogen bonds are more dominant than the backbone hydrogen bonds in the unpaired nucleotides. Arg, Lys, and Ser were commonly found to frequently bind to nucleic acids in our analysis as well as other studies [6,14,15]. These amino acids are hydrophilic and have at least one electronegative atom in their side chains.

4. Discussion

The hydrogen bonding interactions between protein and RNA were analyzed to determine the main features in terms of the interaction propensity on the surface, the atomic level properties in the binding sites, the relationship between the main or side chain and the base or backbone hydrogen bonds, and the relationship between the secondary structures.

Table 8
Distribution of hydrogen bonds at the RNA secondary structure level

Amino acid	Nucleotide								Total
	A		G		C		U		
	P	NP	P	NP	P	NP	P	NP	
Arg	18	36	37	26	51	51	33	54	306
Lys	10	90	25	25	27	31	21	28	257
Ser	8	40	30	21	14	22	6	23	164
Thr	11	65	7	37	8	1	14	8	151
Glu	2	10	19	77	9	12	5	2	136
Asn	6	11	18	14	5	25	30	16	125
Asp	2	0	16	52	9	28	6	3	116
Gln	0	4	20	4	7	7	5	14	61
Tyr	1	12	6	4	3	18	6	9	59
Gly	4	2	18	0	11	2	0	3	40
His	2	6	10	5	4	0	4	5	36
Phe	0	0	0	31	0	0	0	0	31
Leu	0	3	7	0	3	4	0	2	19
Ala	0	3	3	1	1	2	6	1	17
Pro	3	0	0	0	2	4	3	0	12
Trp	0	5	4	0	3	0	0	0	12
Ile	2	4	2	0	2	0	0	0	10
Met	1	2	2	2	1	0	0	1	9
Cys	0	4	0	0	0	0	0	0	4
Val	0	0	0	0	2	1	0	0	3
Total	70	297	224	299	162	208	139	169	1568

P: paired nucleotide, NP: unpaired nucleotide.

The interaction propensity function for this analysis indicated the frequency of the co-occurrences of the amino acids and nucleotides in the protein–RNA complexes for every combination of amino acids and nucleotides. This interaction propensity function is a more refined one than the others since the primary focus in this study was the RNA and the protein. This study found that polar and charged residues, such as Arg, Lys and Thr, showed a high propensity, while the buried and hydrophobic residues, Cys, Val, and Met, had a low propensity. Among the four nucleotides with similar propensity values, uracil showed the highest propensity. High interaction propensities were observed in the Arg–U, Thr–A, Lys–A, and Asn–U pairs. Among the hydrophobic residues, the Trp–C, Trp–U, and Phe–G pairs showed a relatively high propensity.

Hydrogen bonds with side chains were more dominant in amino acids than those with main chains, while there is little difference in the base and backbone hydrogen bonds of the nucleotides. In some amino acids, the side chains are long and have many electronegative atoms that help form hydrogen bonds. All the aliphatic residues, such as Ala, Gly, Ile, Leu, and Val, and the two cyclic residues, Phe and Pro, had no side chain hydrogen bonds. The hydrophobic residues generally favor the main chain hydrogen bonds, except Trp and Tyr. All the aromatic residues, such as Trp, His, and Tyr, prefer the side chain hydrogen bonds to the main chain hydrogen bonds.

The backbone hydrogen bonds were more dominant than the base hydrogen bonds in the paired nucleotides, but these preferences were reversed in the unpaired nucleotides. While paired bases were much more abundant (70%) in the RNA structure than unpaired bases (30%), over 60% of hydrogen bonds were observed in the unpaired bases of RNAs. Especially, the β-sheet residues in proteins have a high tendency to recognize the unpaired nucleotides. The distribution of hydrogen bonds clearly demonstrates that specific nucleotide–protein secondary structures were favored in protein–RNA interactions. For example, a high binding tendency between adenine and sheet residues, cytosine and helix residues, was revealed. For the particular pairings, Lys and the unpaired adenine as well as Glu along with the unpaired guanine provide the greatest specificity. The interaction patterns discovered from the analysis will provide us with useful information

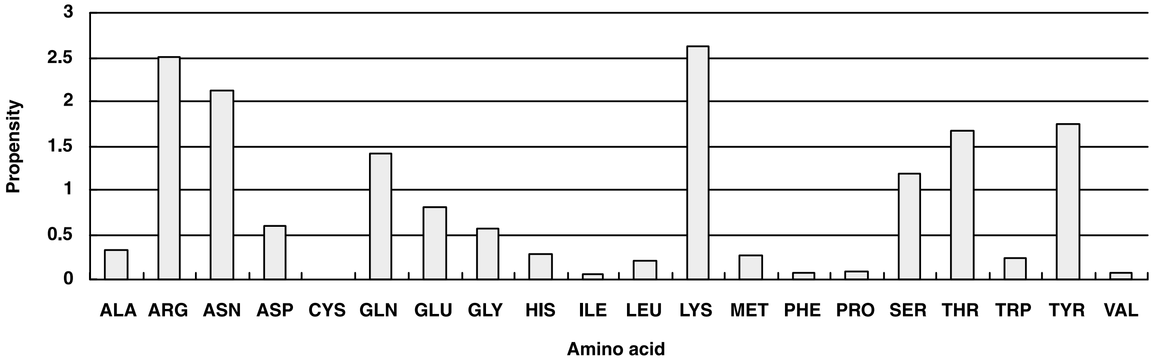

Fig. 5. Average interaction propensity values of the amino acids in Jones' data set, computed by our interaction propensity function.

in predicting the structure of the RNA binding protein and the structure of the protein binding RNA.

Acknowledgements: This work was supported by the Ministry of Information and Communication of Korea under Grant 01-PJ11-PG9-01BT00B-0012.

References

[1] Kim, H., Jeong, E., Lee, S.W. and Han, K. (2002) Genome Inform. 13, 312–313.

[2] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucleic Acids Res. 28, 235–242.

[3] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.

[4] McDonald, I.K. and Thornton, J.M. (1994) J. Mol. Biol. 238, 777–793.

[5] Moodie, S.L., Mitchell, J. and Thornton, J.M. (1996) J. Mol. Biol. 263, 486–500.

[6] Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M. and Thornton, J.M. (2001) Nucleic Acids Res. 29, 943–954.

[7] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.

[8] Nagaswamy, U., Voss, N., Zhang, Z. and Fox, G.E. (1983) Nucleic Acids Res. 28, 375–376.

[9] Cheng, A.C., Chen, W.W., Fuhrmann, C.N. and Frankel, A.D. (2003) J. Mol. Biol. 327, 781–796.

[10] Draper, D.E. (1999) J. Mol. Biol. 293, 255–270.

[11] Varani, G. (1997) Acc. Chem. Res. 30, 189–195.

[12] Nobeli, I., Laskowski, R.A., Valdar, W.S.J. and Thornton, J.M. (2001) Nucleic Acids Res. 29, 4294–4309.

[13] Treger, M. and Westhof, E. (2001) J. Mol. Recogn. 14, 199–214.

[14] Allers, J. and Shamoo, Y. (2001) J. Mol. Biol. 311, 75–86.

[15] Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. (2001) Nucleic Acids Res. 29, 2860–2874.