FEBS Letters 581 (2007) 1881–1890

PRI-Modeler: Extracting RNA structural elements from PDB files of protein–RNA complexes

Kyungsook Han*, Chirag Nepal

School of Computer Science and Engineering, Inha University, Inchon 402-751, Republic of Korea

Received 20 February 2007; accepted 29 March 2007

Available online 9 April 2007

Edited by Ivan Sadowski

Abstract **A complete understanding of protein and RNA structures and their interactions is important for determining the binding sites in protein–RNA complexes. Computational approaches exist for identifying secondary structural elements in proteins from atomic coordinates. However, similar methods have not been developed for RNA, due in part to the very limited structural data so far available. We have developed a set of algorithms for extracting and visualizing secondary and tertiary structures of RNA and for analyzing protein–RNA complexes. These algorithms have been implemented in a web-based program called PRI-Modeler (protein–RNA interaction modeler). Given one or more protein data bank files of protein–RNA complexes, PRI-Modeler analyzes the conformation of the RNA, calculates the hydrogen bond (H bond) and van der Waals interactions between amino acids and nucleotides, extracts secondary and tertiary RNA structure elements, and identifies the patterns of interactions between the proteins and RNAs. This paper presents PRI-Modeler and its application to the hydrogen bond and van der Waals interactions in the most representative set of protein–RNA complexes. The analysis reveals several interesting interaction patterns at various levels. The information provided by PRI-Modeler should prove useful for determining the binding sites in protein–RNA complexes. PRI-Modeler is accessible at http://wilab.inha.ac.kr/primodeler/, and supplementary materials are available in the analysis results section at http://wilab.inha.ac.kr/primodeler/.**

Keywords: Protein–RNA interaction; Hydrogen bond; van der Waals interaction; Visualization of RNA structure

1. Introduction

Most of the currently known structures of molecules were determined by X-ray crystallography or nuclear magnetic resonance (NMR). These methods generate a large amount of structure data consisting mostly of the three-dimensional coordinates of the atoms, even for a small molecule. The coordinate values at the atomic level provide useful information for analyzing molecular structures, but structure elements at higher levels are also required for a complete understanding of structures and, in particular, for predicting structures. There exist computational approaches for assigning secondary structural elements to proteins from their atomic coordinates [1], but similar methods have not been developed for RNA, due to the very small amount of structure data so far available for RNA. Therefore, extracting secondary or tertiary structural elements of RNA requires a substantial amount of manual work. As the number of three-dimensional structures of RNA molecules is increasing, we need a more systematic and automated method for extracting the structural elements of RNA.

We have developed a set of algorithms and a web-based program called PRI-Modeler (http://wilab.inha.ac.kr/primodeler/), which recognizes the secondary and tertiary structures of RNA, and analyzes the hydrogen bond and van der Waals interactions between protein and RNA from the three-dimensional atomic coordinates of protein–RNA complexes in the protein data bank [2]. PRI-Modeler identifies base pairs and classifies base pairs into 28 types [3] to extract secondary and tertiary structures. The secondary and tertiary structures derived from the base pairs are then presented for scrutiny. One advantage of PRI-Modeler is that it can analyze multiple PDB files at once and provides the analysis report on protein–RNA interactions in all the PDB files analyzed. To our knowledge this is the first attempt to extract RNA structural elements from the atomic coordinates in structure databases. Although PRI-Modeler is designed for analyzing RNA structures in protein–RNA complexes, it can also be used for analyzing the interactions between DNA and protein in protein–DNA complexes.

2. Materials and methods

2.1. Base pairs and base-pair rules

An RNA nucleotide consists of a sugar molecule, a molecule of phosphoric acid, and a base. A base pair is formed when one base is paired with another base by hydrogen bonds (H bonds). Base pairs can be classified into canonical base pairs (Watson–Crick base pairs) and non-canonical base pairs. We consider base pairs of 28 types, comprising both canonical and non-canonical base pairs [3]. Supplementary Fig. 1 shows two base pairs.

A base consists of a fixed number of atoms (see Supplementary Fig. 1) that provide important clues for extracting base pairs and classifying them. Base pairs are formed by hydrogen bonding between the atoms of bases. For example, the canonical A–U pair has two H bonds between the N1 of adenine (A) and the N3 of uracil (U), and between the N6 of A and the O4 of U. Thus, we can establish definite rules for H bonding between fixed atoms, which we call base-pair rules. We extract the base pairs formed by H bonds, and classify the base pairs into 28 types using the base-pair rules. The base-pair rules are also used for extracting secondary and tertiary structural elements of RNA.

*Corresponding author. Fax: +82 32 863 4386.
E-mail addresses: khan@inha.ac.kr (K. Han), chirag@inhaian.net (C. Nepal).

0014-5793/$32.00
doi:10.1016/j.febslet.2007.03.085

2.2. H bonds and van der Waals interactions

The number of H bonds between amino acids and nucleotides in protein–RNA complexes was calculated using CLEAN, a program for tidying Brookhaven files, and HBPLUS version 3.15 [4], a program to calculate the number of H bonds. The positions of hydrogen atoms (H) were inferred from the surrounding atoms since hydrogen atoms are invisible in purely X-ray-derived structures. H bonds were identified by finding all proximal atom pairs between H bond donors (D) and acceptors (A) that satisfy the following geometric criteria: contacts with D–A distance <3.35 Å, H–A distance <2.7 Å, and both D–H–A and H–A–AA angles $\geqslant 90°$, where AA is an acceptor antecedent. All protein–RNA bonds were extracted from the HBPLUS output files.

The criteria considered for the van der Waals interactions were: contacts with D–A distance in the range 3.35–3.9 Å, H–A distance $\leqslant 7$ Å, and both D–H–A and H–A–AA angles $\geqslant 90°$. The interactions that were classified as H bonds are not included in the van der Waals interactions. The criteria for the H bonds are similar to those used in the study by Jones et al. [5], but are slightly different from those used in the study of Treger and Westhof [8]. Our criteria for the H bonds are more stringent than those in the study of Allers and Shamoo [15].

2.3. Dataset of protein–RNA complexes

The protein–RNA complex structures were obtained from the protein data bank (PDB) [2]. Complexes determined by X-ray crystallography with a resolution of 3.0 Å or better were selected. As of August 2006, there were 479 protein–RNA complexes in PDB and the number of complexes with a resolution of 3.0 Å or better was 250. We used PSI-BLAST [6] for a similarity search on each of the protein and RNA sequences in these 250 protein–RNA complexes in order to eliminate equivalent amino acids or nucleotides in homologous protein or RNA structures. Forty-five of the protein–RNA complexes remained as representative, non-homologous complexes after running the PSI-BLAST program with an E value of 0.001 and an identity value of 90% or below. Thus the final dataset consisted of 45 protein–RNA complexes (Table 1).

2.4. Secondary structure elements

The secondary structures of proteins are assigned using the DSSP program [1]. The secondary structure data of a protein are needed to generate the interaction patterns between the protein and RNA at the secondary structure level. Given the atomic coordinates of a protein in the PDB format, DSSP defines its secondary structure elements, geometrical features and solvent exposure. The secondary structure elements defined by DSSP are classified into four types: helix (α-helix, 3/10 helix and π-helix), sheet (β-ladder and β-bridge), turn (hydrogen-bonded turn), and other (bend and other structures). The RNA–RNA interactions extracted from the HBPLUS output are used to assign a secondary structure to each nucleotide.

We consider two types of RNA secondary structure elements: paired and unpaired. If at least one H bond exists between the base parts of two nucleotides, these nucleotides are considered to be paired. If not, they are considered to be unpaired. Two nucleotides with canonical base pairs, wobble base pairs, and all non-canonical pairs reported by Nagaswamy et al. [7] are classified as paired nucleotides.

2.5. Classification of amino acids and nucleotide atoms

An amino acid can exhibit various properties, and so can be classified into two or more categories. We classified the amino acids into six different types based on their physico-chemical properties (http://www.russell.embl-heidelberg.de/aas/aas.html): hydrophobic (Val, Ile, Leu, Met, Phe, Trp, and Cys), polar (Arg, Asp, Asn, Glu, Gln, and Lys), aromatic (His, Phe, Trp, and Tyr), aliphatic (Ala, Pro, Ile, Leu, and Val), positively charged (Arg and Lys) and negatively charged (Glu and Asp).

The atoms of nucleotides are conventionally grouped into three parts [8]. We consider that atoms C1*, C2*, C3*, C4*, C5*, O2*, and O4* belong to the ribose, and atoms P, O1P, O2P, O3P, O3*, and O5* to the phosphate. The remaining atoms belong to the bases.

2.6. Interaction propensity

Counting the number of H bonds does not yield precise interaction propensities because it does not take into account the number of residues in the complex and on the surface of the complex. For example, amino acid A may have a weak interaction propensity even though it is involved in many H bonds or occurs very frequently. Therefore, we employ a function for determining interaction propensity. It is based on the function employed by Moodie et al. [9], but has been modified to generate the interaction propensity of every combination of amino acids and nucleotides on the surface of a complex. Amino acids are considered to be on the surface if their relative accessibility exceeds 5% according to the Naccess program [10].

The interaction propensity P_{ab} between amino acid a and nucleotide b is defined by Eq. (1), where N_{ab} is the number of amino acids a hydrogen bonded to nucleotide b, $\sum N_{ij}$ is the total number of amino acids hydrogen bonded to any nucleotide, N_a is the number of amino acids a, $\sum N_i$ is the total number of amino acids, N_b is the number of nucleotides b, $\sum N_j$ is the total number of nucleotides, and the numbers refer to residues on the surface. The numerator $N_{ab}/\sum N_{ij}$ represents the ratio of the co-occurrences of amino acid a binding with nucleotide b to the total number of amino acids binding to any nucleotide on the surface. The term $N_a/\sum N_i$ in 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 $N_b/\sum N_j$ represents the ratio of the frequency of nucleotide b to that of all nucleotides on the surface

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

It should be noted that the interaction propensity of Eq. (1) is calculated as the proportion of a particular amino acid binding to a particular nucleotide on the surface divided by the proportion of each on the surface. Therefore, the propensity value represents the frequency of co-occurrence of amino acids and nucleotides in protein–RNA complexes for each combination of amino acid and nucleotide. A propensity greater than 1 indicates that co-occurrence of the given amino acid and nucleotide is high, whereas a propensity less than 1 indicates that cooccurrence of the given amino acid and nucleotide is low on the surface.

2.7. Algorithm

PRI-Modeler consists of a set of algorithms. Fig. 1 shows the framework for extracting the information about the secondary and tertiary structures of RNA. We use HBPLUS [4] to get all H bonds in a PDB file. Algorithm 1 extracts the RNA sequence data from a PDB file and records it in RNA-SEQ.

Algorithm 2 extracts only H bonds between a base of a nucleotide and a base from the H bonds obtained by HBPLUS. These H bonds between pairs of bases are recorded in a Base–Base List, and classified into the 28 types. The atoms of a base are numbered, and H bonds are formed between fixed atoms for each base pair. These atom numbers are used when determining base pairs. The H bond data in the Base–Base List are analyzed by Algorithm 2, and the base pairs extracted are recorded in the Base-Pair List with their types.

Table 1
The list of 45 protein–RNA complexes in the dataset

PDB IDs									
1A9N	1AQ3	1ASY	1B7F	1B23	1C9S	1DFU	1DI2	1DK1	1DUL
1E7K	1EC6	1EUY	1F7U	1F8V	1FFY	1G59	1H3E	1H4Q	1HC8
1I6U	1JBR	1JID	1K8W	1M8Y	1MJI	1MZP	1R3E	1RC7	1RPU
1SI3	1TFW	1TTT	1WSU	1XOK	1YTY	1YVP	2ASB	2AZ0	2B3J
2BGG	2BTE	2BX2	2F8K	2FMT					

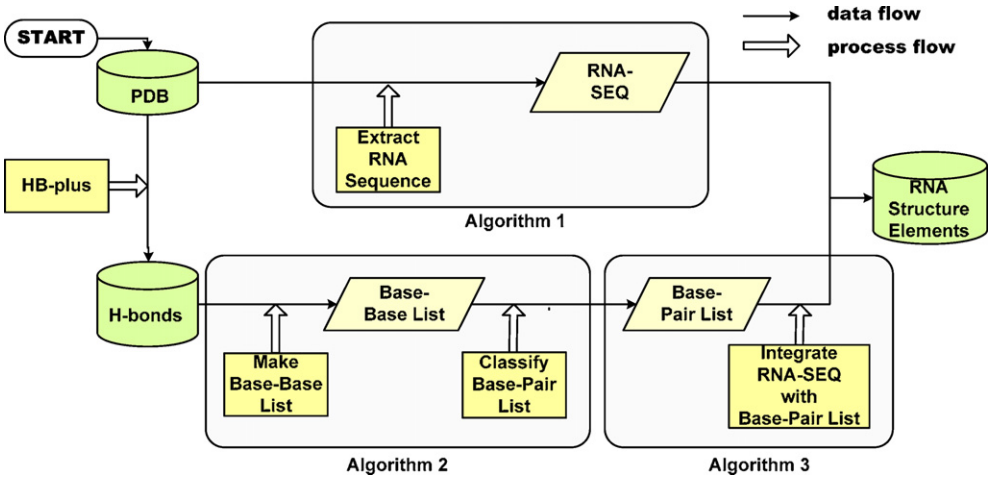

Fig. 1. Framework for extracting base pairs of 28 types and secondary and tertiary structure elements of RNA from PDB files.

Algorithm 3 integrates the RNA sequence data obtained by Algorithm 1 with the base-pair data. During this process the algorithm analyzes all the nucleotides of the RNA in RNA-SEQ and matches the nucleotides in the Base-Pair List to the nucleotides in RNA-SEQ to find the H-bonding relation of each nucleotide. The final output of the three algorithms is the secondary and tertiary structure elements of the RNA in the protein–RNA complex. The algorithms are outlined below.

Algorithm 1

```
Given a PDB file P
for all atoms a in P do
   if a is nucleotide of RNA then
      RNA-SEQ = RNA-SEQ ∪ a
   end if
end for
```

Algorithm 2

```
Given H bond data in H bonds
for all H-bonds b in H bonds do
   if b is an B-bond between RNA nucleotides then
      b is bb
      if bb is an H bond between bases then
       Base-Base-List = Base-Base-List ∪ bb
       end if
      end if
   end for
   Let H bonds between bases in the Base-Base-List bb-
   bonds
   for all bb-bond in Base-Base-List do
      if a base-pair is constructed by the bb-bond then
       Base-Pair-List = Base-Pair-List ∪ bb-bond
      end if
   end for
```

Algorithm 3

```
for all nucleotide n in RNA-SEQ do
   for all bp-bond in the Base Pair Lists do
      if n is assigned bp-bond then
            paired nucleotide = n ∪ bp-bond
      else
       unpaired nucleotide = n
      end if
      RNA-structure-List = RNA-structure-
      List ∪ Paired nucleotide ∪ unpaired nucleotide
   end for
end for
```

PRI-Modeler also derives the three-dimensional coordinates of a nucleotide by averaging the coordinate values of all the atoms that make up the nucleotide. The three dimensional coordinates of the nucleotides are integrated with the structure elements of RNA to visualize the structure.

3. Results

3.1. Implementation of PRI-Modeler

PRI-Modeler was developed using Microsoft Visual C#, and is executable within a web browser on any PC with Windows 2000/XP/Me/98/NT 4.0 as its operating system. Given one or more PDB files as input, PRI-Modeler produces several files containing H bonds and van der Waals interactions between amino acids and nucleotides, secondary and tertiary structures of RNA, and interaction patterns of protein–RNA complexes. It also visualizes the RNA structures extracted from protein–RNA complexes with H bonds and van der Waals interactions highlighted. PRI-Modeler is accessible at http://wilab.inha.ac.kr/primodeler/default.htm, and the supplementary material is available in the analysis results section at http://wilab.inha.ac.kr/protein_RNA/.

Fig. 2 shows an exemplary user interface of PRI-Modeler. It consists of 20 menu buttons and one text panel. When the user specifies the types of interaction to analyze using the menu buttons (such as H bonding interactions and van der Waals interactions), PRI-Modeler analyzes the specified interactions and displays the results in the text panel. The remainder of this paper presents the results of an analysis of the 45 protein–RNA complexes referred in Section 2.

3.2. Extraction of secondary and tertiary RNA structures

PRI-Modeler can extract structure elements formed by multiple chains. Fig. 3A depicts an RNA structure formed by four RNA chains, together with the H bonds between the nucleotides, extracted from a protein–RNA complex (PDB ID: 1FEU). Each node represents a nucleotide, and the blue lines and red dotted lines represent RNA backbones and H bonds between nucleotides, respectively. The nucleotides are numbered at intervals of 5 nucleotides. One can rotate the structure as well as zoom in/out to get a clear perspective view.

Since PRI-Modeler uses base pair information to identify secondary and tertiary RNA structures, it can also easily retrieve base triplets. A base triplet is a tertiary RNA interaction

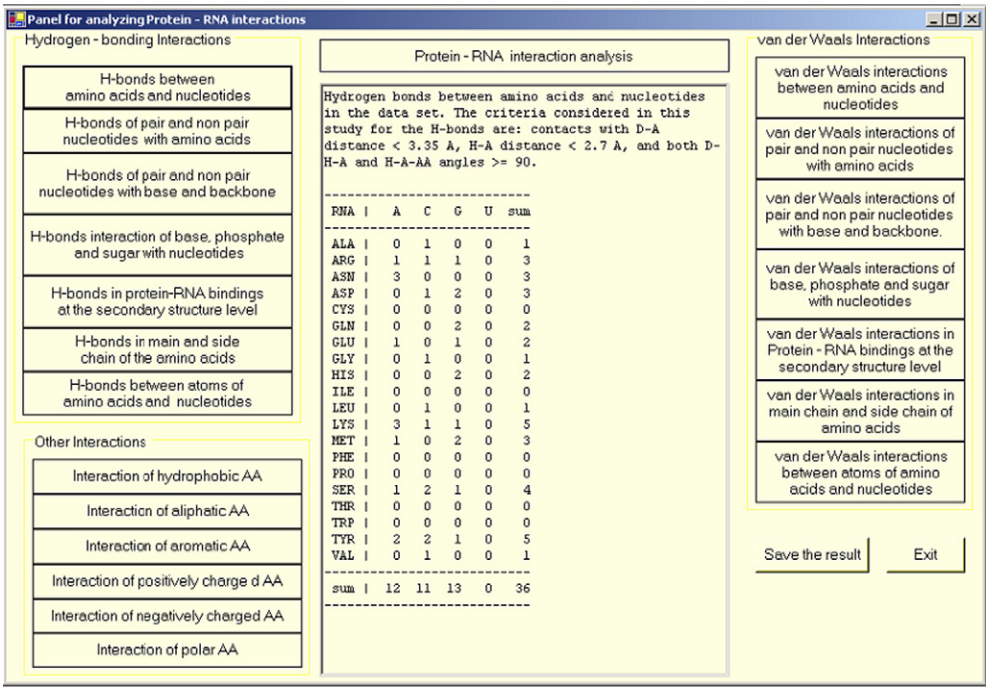

Fig. 2. User interface of PRI-Modeler for analyzing the interactions between protein and RNA.

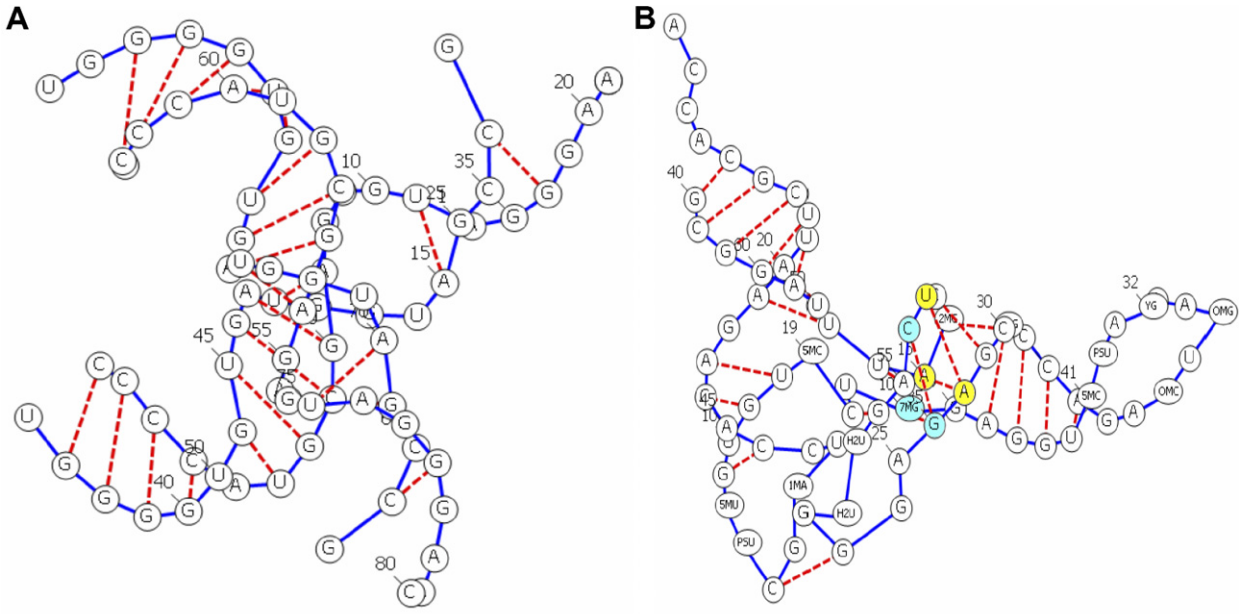

Fig. 3. (A) RNA structure with four chains, extracted from a protein–RNA complex (PDB ID: 1FEU) with H bonds in red dotted lines. (B) RNA structure (PDB ID: 1EHZ) obtained by PRI-Modeler. The yellow nodes connected by red dotted lines and cyan nodes connected by red dotted lines are base-triplets.

in which a pair of bases interact with a third base [11]. For example, in the C–G–G base triplet, the G–C pair is a secondary interaction involving a canonical pair, and the G–G pair is a tertiary interaction involving a purine–purine pair (see Supplementary Fig. 2 for the base triplet structure). The base triplets U–A–A and C–G–G shown in Fig. 3B were extracted by PRI-Modeler from the structure of yeast phenylalanine tRNA (PDB ID: 1EHZ), and agree with those found experimentally [12].

Programs like RASMOL (http://www.umass.edu/microbio/rasmol/) and MOLSCRIPT [13] can generate a drawing of the structure of a molecule from the three-dimensional coordinates of its atoms. However with such programs one cannot easily obtain information about each nucleotide in the RNA,

or the bonds between nucleotides, because these programs represent molecular structures at the atomic level. PRI-Modeler uses the three-dimensional coordinates of nucleotides to generate stereoscopic images of the secondary and tertiary structures of RNA. Furthermore, it derives the configuration of a given RNA molecule as well as the bonding relations and types of base pairing between nucleotides.

3.3. Amino acids in contact with RNA

An amino acid is considered to interact with RNA if an atom of the amino acid interacts with an atom of the RNA. Tables 2 and 3 show H bonding and van der Waals interactions, respectively. There are a total of 1442 H bonds and 754 van der Waals contacts between amino acids and nucleotides. The amino acids in the first column are sorted in

Table 2
Hydrogen bonds between amino acids and nucleotides and the interaction propensity

Amino acids	Frequency	Nucleotides					Interaction propensity (IP)				Average IP
		A	C	G	U	Total	A	C	G	U	
Glu	3390	13	22	**51**	6	92	0.42	0.45	1.14	0.20	0.55
Lys	2866	**54**	**70**	**71**	**41**	**236**	**2.06**	**1.69**	**1.88**	1.66	**1.82**
Arg	2728	**67**	**99**	**131**	**86**	**383**	**2.69**	**2.51**	**3.64**	**3.67**	**3.12**
Leu	2270	5	8	0	3	16	0.24	0.24	0.00	0.15	0.15
Gly	2096	9	16	19	6	50	0.47	0.53	0.68	0.33	0.50
Asp	1960	10	14	40	16	80	0.77	0.49	1.54	0.95	0.93
Thr	1838	20	18	32	29	99	1.19	0.68	1.32	1.83	1.25
Ala	1764	2	9	6	10	27	0.12	0.35	0.25	0.66	0.34
Val	1688	1	3	0	0	4	0.06	0.12	0.00	0.00	0.04
Ser	1640	**32**	**33**	24	13	**102**	**2.13**	1.40	1.11	0.92	1.39
Pro	1628	4	3	2	0	9	0.27	0.12	0.09	0.00	0.12
Asn	1416	21	16	30	**36**	**103**	1.62	0.78	1.60	**2.96**	**1.74**
Gln	1290	16	28	32	22	98	1.35	1.50	**1.88**	1.98	**1.67**
lle	1206	10	2	0	5	17	0.91	0.11	0.00	0.48	0.37
Tyr	1122	9	23	4	21	57	0.88	**1.42**	0.27	**2.18**	1.18
Phe	1032	0	0	21	3	24	0.00	0.00	1.54	0.34	0.47
His	870	6	12	9	3	30	0.75	0.95	0.78	0.40	0.72
Met	474	2	2	4	0	8	0.46	0.29	0.64	0.00	0.34
Trp	382	0	4	1	1	6	0.00	0.72	0.20	0.30	0.30
Cys	172	1	0	0	0	1	0.63	0.00	0.00	0.00	0.15
Total	31 832	282	382	477	301	1442					
Average IP							0.85	0.71	0.92	0.95	

Table 3
van der Waals interactions between amino acids and nucleotides and the interaction propensity

Amino acids	Frequency	Nucleotides					Interaction propensity (IP)				Average
		A	C	G	U	Total	A	C	G	U	
Glu	3390	5	10	12	5	32	0.30	0.39	0.51	0.32	0.38
Lys	2866	16	**28**	**19**	**32**	**95**	1.17	1.30	0.96	**2.48**	**1.47**
Arg	2728	**32**	**71**	**78**	**63**	**244**	**2.45**	**3.45**	**4.14**	**5.14**	**3.79**
Leu	2270	1	3	1	1	6	0.09	0.17	0.06	0.10	0.10
Gly	2096	4	10	14	**11**	39	0.40	0.63	0.97	1.17	0.79
Asp	1960	4	5	5	3	17	0.42	0.33	0.37	0.34	0.36
Thr	1838	**17**	16	18	4	55	**1.93**	1.15	1.42	0.48	1.24
Ala	1764	1	3	2	1	7	0.11	0.22	0.16	0.12	0.15
Val	1688	0	0	0	1	1	0.00	0.00	0.00	0.19	0.04
Ser	1640	**38**	17	6	8	**69**	**4.85**	1.37	0.53	1.08	**1.95**
Pro	1628	0	4	0	1	5	0.00	0.32	0.00	0.13	0.11
Asn	1416	9	9	**31**	**11**	**60**	1.33	0.84	**3.17**	**1.73**	**1.76**
Gln	1290	8	11	15	7	41	1.30	1.13	1.68	1.20	1.32
Ile	1206	1	1	0	0	2	0.17	0.11	0.00	0.00	0.07
Tyr	1122	4	**22**	7	6	39	0.74	**2.60**	0.90	1.19	1.35
Phe	1032	3	0	3	1	7	0.61	0.00	0.42	0.21	0.31
His	870	1	6	11	0	18	0.24	0.91	1.83	0.00	0.74
Met	474	2	2	1	1	6	0.88	0.56	0.30	0.47	0.55
Trp	382	0	4	0	1	5	0.00	**1.39**	0.00	0.58	0.49
Cys	172	1	1	4	0	6	1.21	0.77	**3.37**	0.00	1.33
Total	31 832	147	223	227	157	754					
Average IP							0.91	0.88	1.03	0.84	

decreasing order of their frequency. Adenine (A), cytosine (C), guanine (G) and uracil (U) form 282, 382, 477 and 301 H bonds, respectively. A, C, G and U form 147, 223, 227 and 157 van der Waals interactions, respectively. Interestingly, in both the H bonding and van der Waals interactions, arginine (Arg) shows the highest interaction propensity, and valine (Val) shows the lowest interaction propensity. The high interaction propensity of Arg can be explained by its unique capacity to form multiple H bonds and to bind to multiple nucleotides simultaneously.

Amino acids are more diverse than nucleotides, with H bonding interaction propensities in the range [0.04–3.12], and van der Waals interaction propensities in the range [0.04–3.79]. In contrast, nucleotides have H bonding interaction propensities in the range [0.71–0.95], and van der Waals interaction propensities in the range [0.84–1.03]. This indicates that amino acids are more distinguishable by their interaction propensities than are nucleotides. Amino acids have a main chain in common, and nucleotides have a backbone in common. In proteins, both H bonding and van der Waals interactions are more frequent in the side chains (average 75%) than in the main chains (average 25%) (Table 4). Amino acids, in which side chain contacts predominate, naturally have more diverse interaction propensities than nucleotides. In contrast, in RNA, the interactions observed in the backbone (on average 67% of all interactions) exceed those in the bases (33%) (Table 5). The backbone part of a nucleotide has more atoms, including electronegative atoms, than the base part, which makes the backbone more favorable for interacting with amino acids than the base.

Among amino acids Arg has the highest H bonding interaction propensity value of 3.12, and Val has the smallest value of 0.04 (Table 2). Arg has a large side chain with many electronegative atoms, and displays diverse interaction patterns. It forms a large number of H bonds and van der Waals interactions with all nucleotides. Arg–G forms the largest number of H bonds (131 H bonds), whereas Arg–U has the highest propensity value of 3.67. Among nucleotides uracil has the highest H bonding interaction propensity value of 0.95 and cytosine the lowest of 0.71. Guanine, the most frequent nucleotide in H bonding interactions with amino acids, forms 477 H bonds (33% of the total 1442 H bonds). Guanine prefers Arg, Lys, and Glu, cytosine prefers Arg, Lys, and Ser, adenine prefers Arg, Lys and Ser, and uracil prefers Arg, Lys and Asn.

In van der Waals interactions, most hydrophobic residues (Val, Ile, Leu, Met, Phe and Trp) have low average interaction propensities, and many polar residues (Arg, Asn, Gln, and Lys) have high average interaction propensities (Table 3). Several amino acids show a strong preference for specific nucleotides with which they form stable van der Waals interactions. Arg has the largest number (244) of van der Waals interactions with nucleotides, and the highest average propensity of 3.79.

Table 4
H bonds and van der Waals interactions in the main chains and side chains of amino acids

Amino acids	H bonds			van der Waals interactions		
	Main chain	Side chain	Total	Main chain	Side chain	Total
Arg	**41**	**342**	383	10	**234**	244
Lys	53	**183**	236	12	**83**	95
Ser	18	**84**	102	**17**	**52**	69
Thr	28	71	99	**17**	38	55
Glu	12	80	92	9	23	32
Asp	17	63	80	8	9	17
Tyr	12	45	57	7	32	39
Gln	7	91	98	3	38	41
Asn	22	81	103	9	51	60
Gly	**50**	0	50	**39**	0	39
His	10	20	30	2	16	18
Ala	27	0	27	7	0	7
Pro	9	0	9	5	0	5
lle	17	0	17	2	0	2
Leu	16	0	16	6	0	6
Val	4	0	4	1	0	1
Met	5	3	8	4	2	6
Trp	3	3	6	1	4	5
Cys	1	0	1	1	5	6
Phe	24	0	24	7	0	7
Total	376	1066	1442	167	587	754

Table 5
H bonds and van der Waals interactions in the bases, phosphates and sugars of nucleotides

	H bonds					van der Waals interactions				
	A	C	G	U	Total	A	C	G	U	Total
Base	108	115	**215**	126	564	23	37	**68**	37	165
Phosphate	87	**139**	122	100	448	28	**65**	50	30	173
Sugar	87	**128**	140	75	430	96	**121**	109	90	416
Total	282	382	477	301	1442	147	223	227	157	754

Lys, Asn, Thr, Ser and Gln are also frequently-bonding amino acids. The frequent nucleotide-amino acid pairs in van der Waals interactions are A (Ser, Arg, Lys), C (Arg, Lys, Tyr), G (Arg, Asn, Lys) and U (Arg, Lys, Gly and Asn). Arg–G pair has the largest number of van der Waals interactions (78), whereas Arg–U pair has the highest interaction propensity of 5.14. Val has the lowest average interaction propensity of 0.04. The van der Waals interaction propensities of nucleotides range from 0.84 (uracil) to 1.03 (guanine).

3.4. Interactions of the main chains and side chains of proteins

Table 4 presents an analysis of H bonds and van der Waals interactions formed by the main chains and side chains of proteins. The main chains have 376 H bonds and side chains have 1066 H bonds. Evidently the side chains have more H bonds (74%) than the main chains (26%). The aliphatic amino acids (Ala, Pro, He, Leu, and Val) have no side chains, and therefore no side chain H bonds, whereas they form 19% of the main chain H bonds (73 out of 376). Gly has a large number of main chain H bonds, but has no side chain H bond. Some aromatic amino acids (His and Tyr) prefer side chain H bonds to main chain H bonds. The charged and polar amino acids (Arg, Lys, Asn, Glu, Gin, Asp) have more side chain interactions than main chain interactions in both H bonds and van der Waals contacts.

78% (587 out of 754) of the van der Waals interactions were observed in side chains and 22% (167 out of 754) in main chains (Table 4). All aliphatic amino acids (Ala, Pro, Ile, Leu, and Val) and many hydrophobic amino acids (Val, Ile, Leu, Met, and Phe) had more van der Waals interactions in main chains than in side chains. Many charged and polar amino acids (Arg, Asn, Gln, and Lys) had a strong preference for side chains, and most van der Waals interactions were observed in the side chains of polar amino acids.

Each nucleotide is composed of a base and a backbone, and the backbone consists of a phosphate and a sugar. As shown in Table 5, the bases, phosphates and sugars of nucleotides form 564, 448 and 430 H bonds, respectively. Hence, the backbone parts of nucleotides form a total of 878 (448 + 430) H bonds (61%). The largest number of H bonds in the base part and the backbone part is observed in guanine (215 H bonds in base) and cytosine (139 + 128 = 267 H bonds in backbone), respectively. In van der Waals interactions, the base part of guanine and the backbone part of cytosine are prominent (Table 5). In both H bonding and van der Waals interactions, the backbone parts of nucleotides have more interactions than the base parts.

3.5. Protein–RNA interactions at the secondary structure level

PRI-Modeler analyzes the protein–RNA complexes to identify the interaction patterns at the secondary structure level. The secondary structure elements of RNA are of two types, i.e., paired (P) and unpaired (NP), while the secondary structure elements of proteins are of four types, i.e., helix, sheet, turn and others. Although Treger and Westhof [8] do not consider the turns of protein, turns constitute 12% of H bonds and 11% of van der Waals interactions (Table 6), and we consider them to be secondary structure elements of proteins. Table 6 shows the H bonds and van der Waals interactions at the secondary structure level of proteins and RNAs. Helices, forming 37% of the total H bonds, are the most interacting secondary structures and turns are the least interacting.

Table 7 shows H bonds at the secondary structure level of RNA. In H bond interactions, Asp and Glu show a strong tendency to interact with unpaired nucleotides, while in van der Waals interactions, Arg shows a tendency to interact with unpaired nucleotides. As shown in Table 8, the backbone part of a nucleotide has more interactions with protein than the base part both in H bonding and van der Waals interactions. In both H bonding interactions and van der Waals interactions, the base part of unpaired nucleotides is preferred to that of paired nucleotides. In van der Waals interactions, unpaired nucleotides are preferred to paired nucleotides.

3.6. Protein–RNA interactions at the atomic level

Both the atoms N and O of amino acids are found in most amino acids, and are involved in 26% of H bonding interactions and 22% of van der Waals interactions (Supplementary Tables 1 and 6). However, either N or O (but not both) is not involved in the H bonding interactions of Cys, Tip and Val, and in the van der Waals interactions of Cys, His and Ile. Different amino acids have different groups of atoms involved in the interactions with RNA nucleotides. For example, the NH2 and NH1 of Arg form 169 and 123 H bonds, respectively, representing 76% of the total H bonds with nucleotides formed by Arg (Supplementary Table 1). Similarly, the atoms NH2 and NH1 of Arg form 120 and 83 van der Waals interactions with nucleotides, representing 83% of the total van der Waals interactions with nucleotides formed by Arg (Supplementary Table 6). The groups ND2, OD1, NE2, OE2, NZ, OG, OG1 and OH are responsible for most of the H bonding interactions in Asn, Asp, Gln, Glu, Lys, Ser, Thr, and Tyr, respectively, while the groups ND2, NE2, NZ,OG, OG1 and OH are responsible for most of the van der Waals interactions in Asn, Gln, Lys, Ser, Thr and Tyr, respectively. Groups OE1 and OE2 of Glu are involved in both H bonding and van der Waals interactions. In both H bonding and van der Waals interactions, same atoms in each amino acid (Asn-ND2, Gln-NE2, Lys-NZ, Ser-OG, and Tyr-OH) are responsible for most

Table 6
H bonds and van der Waals interactions between protein and RNA at the secondary structure level of protein

	H bonds									van der Waals interactions								
	A		C		G		U		Total	A		C		G		U		Total
	P	NP	P	NP	P	NP	P	NP		P	NP	P	NP	P	NP	P	NP	
Helix	38	52	83	57	129	65	55	56	535	18	22	33	46	43	61	11	39	273
Sheet	38	68	32	30	36	83	21	59	367	14	28	15	24	19	22	2	51	175
Turn	11	22	34	26	42	12	23	9	179	3	14	16	14	13	7	4	14	85
Others	18	35	48	72	51	59	33	45	361	7	41	29	46	33	29	5	31	221
Total	105	177	197	185	258	219	132	169	1442	42	105	93	130	108	119	22	135	754

Table 7
H bonds and van der Waals interactions between protein and RNA at the secondary structure level of RNA

	H bonds									van der Waals interactions								
	A		C		G		U		Total	A		C		G		U		Total
	P	NP	P	NP	P	NP	P	NP		P	NP	P	NP	P	NP	P	NP	
Ala	0	2	1	8	5	1	4	6	27	0	1	1	2	0	2	0	1	7
Arg	24	43	57	42	67	64	37	49	383	8	24	20	51	43	35	7	56	244
Asn	13	8	11	5	28	2	15	21	103	2	7	4	5	9	22	3	8	60
Asp	2	8	5	9	11	29	4	12	80	0	4	4	1	0	5	1	2	17
Cys	0	1	0	0	0	0	0	0	1	1	0	1	0	3	1	0	0	6
Gln	6	10	16	12	23	9	8	14	98	1	7	6	5	8	7	0	7	41
Glu	6	7	7	15	19	32	0	6	92	1	4	1	9	5	7	0	5	32
Gly	5	4	9	7	16	3	3	3	50	2	2	5	5	10	4	4	7	39
His	4	2	6	6	9	0	0	3	30	1	0	3	3	5	6	0	0	18
Ile	0	10	2	0	0	0	0	5	17	0	1	1	0	0	0	0	0	2
Leu	0	5	2	6	0	0	0	3	16	0	1	3	0	1	0	0	1	6
Lys	15	39	43	27	44	27	24	17	236	7	9	11	17	11	8	1	31	95
Met	0	2	2	0	2	2	0	0	8	0	2	1	1	0	1	0	1	6
Phe	0	0	0	0	0	21	0	3	24	0	3	0	0	1	2	0	1	7
Pro	0	4	1	2	2	0	0	0	9	0	0	4	0	0	0	0	1	5
Ser	17	15	11	22	20	4	7	6	102	13	25	12	5	3	3	3	5	69
Thr	11	9	6	12	11	21	23	6	99	4	13	4	12	9	9	0	4	55
Trp	0	0	4	0	1	0	1	0	6	0	0	2	2	0	0	1	0	5
Tyr	1	8	12	11	0	4	6	15	57	2	2	10	12	0	7	2	4	39
Val	1	0	2	1	0	0	0	0	4	0	0	0	0	0	0	1	0	1
Total	105	177	197	185	258	219	132	169	1442	42	105	93	130	108	119	22	135	754

Amino acids are listed in the alphabetical order.

Table 8
H bonds and van der Waals interactions in the base and backbone parts of nucleotides

	H bonds									van der Waals interactions								
	A		C		G		U		Total	A		C		G		U		Total
P/NP	P	NP	P	NP	P	NP	P	NP		P	NP	P	NP	P	NP	P	NP	
Base	22	**86**	19	**96**	77	**138**	29	**97**	564	6	17	8	**29**	32	36	4	**33**	165
Backbone	83	91	**178**	89	**181**	81	103	72	878	36	**88**	85	**101**	76	83	18	**102**	589
Total	105	177	197	185	258	219	132	169	1442	42	105	93	130	108	119	22	135	754

interactions of the amino acid. Details are shown in Supplementary Tables 1 and 6.

Likewise, different nucleotides use different atoms for their H bonding and van der Waals interactions with amino acids. Some atoms are prominent in all interactions while others participate in some interactions only or in no interaction at all. The atom O2* is involved in 22% of H bonds formed by nucleotides, which is similar to the 21% reported by Nadassy et al. [14]. However, O2* is involved in 26% of the van der Waals interactions in our study, which is slightly greater than 21% in Nadassy et al. [14]. The atoms O2*, N1, N6, O1P and O2P of adenine constitute 87% of the H bonds formed by adenine (Supplementary Table 2). On the other hand, the atoms O2, O2*, and O1P of cytosine (Supplementary Table 3), O2*, N2, O6, O1P and O2P of guanine (Supplementary Table 4), and O2* O4, O1P, and O2P of uracil constitute the majority of the H bonds (Supplementary Table 5). Similarly in the case of van der Waals contacts, the atoms A (O2*, O3*, O1P, O2P), C (O2*, O3*, O1P, O2P), G (O2*, O3*, O1P, O6) and U (O2*, O3*, O1P, O4*) are responsible for the majority of interactions with amino acids (Supplementary Tables 7–10). Both in H bonding and van der Waals interactions, O2*, O1P and O2P of nucleotides are the preferred atoms.

4. Discussion and comparison with other studies

Here we compare our results with previous works by Jones et al. [5], Allers and Shamoo [15] and Treger and Westhof [8]. Direct comparison of our analysis with these studies was not easy because (1) all these used different criteria from each other for hydrogen bond and van der Waals interaction, (2) they analyzed different datasets, and (3) they used slightly different interaction propensity functions, if any.

Our criteria for the H bonds and van der Waals interactions are similar to those used in the study by Jones et al. [5]. Our analysis shows similar results to these studies in some aspects, but does not fully agree with them. The differences are caused by the differences in the criteria for H bond and van der Waals interactions, datasets and the interaction propensity functions. While our interaction propensity function yields the binding propensity of an amino acid for each of the nucleotides, Jones' function does not distinguish between the different nucleotides binding to an amino acid. Their propensity values range from 0.2 to 1.7, whereas ours range from 0 to 3.67 in H bonds and from 0 to 5.14 in van der Waals interactions. In Jones' study, H bonds and van der Waals contacts constitute 8% and 92% of interactions, respectively. In Treger and Weshtof's study H

bonds constitute 23% of the total interactions and van der Waals contacts constitute 72%, whereas in our study H bonds constitute 65% and van der Waals contacts 35% of the total interactions. Table 9 summarizes the comparisons, and more details are stated below.

In our analysis the residues most preferred in H bonding interactions are Arg, Lys, Asn and Ser (in decreasing order of preferences), and the least preferred residues are Cys, Val, Trp, and Met. In Jones' analysis the most preferred residues are Arg, Asn, Thr, Tyr and Lys, and the least preferred are Cys, Ile, Met, Phe and Pro. In our analysis the residues most preferred in van der Waals interactions are Arg, Lys, Ser and Asn. In Jones' analysis the most preferred residues in van der Waals interactions are Arg, Tyr, Phe, Asn and Thr. In our analysis, as in that of Jones et al., guanine is the most preferred nucleotide. The frequent pairs of H bonding in our analysis are A (Arg, Lys, Ser), C (Arg, Lys, Ser), A (Arg, Lys, Glu), and A (Arg, Lys, Asn), whereas the frequent pairs in Jones' analysis are A (Ser, Phe), C (Asp, Tyr), G (Trp, Gly), and U (Asn, Glu). While our analysis shows that the backbone part is preferred for both H bonding and van der Waals interactions, Jones' data indicate no preference between the base and backbone. However in both our analysis and theirs, side chains are more prominent than main chains in H bonding and van der Waals interactions, and the sheet structures prefer to bind with unpaired RNA nucleotides.

In the analysis of Allers and Shamoo [15], the atoms Nl and N6 constitute 73% of the total H bonds formed by adenine, whereas they make up only 19% in our study (54 out of a total of 282 H bonds formed by adenine). The atoms O2*, Nl, N6, O1P and O2P are responsible for most H bonding interactions formed by adenine in our study (Supplementary Table 2). In Allers and Shamoo's study, the atoms O2, N3, and N4 are the most frequent interacting atoms of cytosine, whereas they are O2* and O1P in our analysis (Supplementary Table 3). The O6 of guanine constitutes 41% of H bonds in Allers and Shamoo's analysis, whereas it forms only 13% (61 out of total of 477 H bonds Supplementary Table 4) in our analysis. In guanine, O2* is the most prominent in H bonding interactions with amino acids, followed by N2, O1P, O6 and O1P in our study. In addition, the atoms O2, N3, and O4 of uracil, the most interacting in Allers and Shamoo's analysis, differ from the atoms O4, O1P and O2* of uracil in our analysis (Supplementary Table 5). In our analysis, the atoms O2P, O2 and N3 of uracil are also frequent in H bonding interactions with amino acids. In general, the atoms O2*, O1P and O2P of each

Table 9
Comparison with other studies on protein–RNA interactions

	Our study	Jones et al. [5]	Allers and Shamoo [15]	Treger and Westhof [8]
Hydrogen bonds (H)				
Proportion of H bonds in the study	65%	8%	–	23%
Frequent amino acids	Arg, Lys, Asn, Ser	Arg, Asn, Thr, Tyr	Arg, Asn, Ser	Arg, Asn, Ser, Lys
Rare amino acids	Cys, Val, Trp, Met	Cys, lle, Met, Phe	Cys, Met, Phe	Ala, lle, Leu, Val
Frequent nucleotide–amino acid pairs	A (Arg, Lys, Ser) C (Arg, Lys, Ser) G (Arg, Lys, Glu) U (Arg, Lys, Asn)	A (Ser, Phe) C (Asp, Tyr) G (Trp, Gly) U (Asn, Glu).	A (Arg, Ser) C (Gln, Ser) G (Arg, Glu) U (Asn)	A (lle , Pro, Ser) C (Leu), G (Asp, Gly) U (Asn)
Nucleotides in the order of preference	G, C, U, A	G, U, C, A	G, C, A, U	No preference
Frequent protein atoms	NH1, NH2, NE2, NZ, ND2, N, O, OG, OG1 OH, OE2,, OD1,	–	–	NH, OH, O, OXT
Frequent RNA atoms	A(O2*,O1P, O2P, N1), C(O2,O2*,O1P), G(O2*,O6, N2, O1P, O2P), U(O2*, O4, O1P, O2P)	O2*	A(N1,N6) C (O2, N3, N4) G(O6) U (O2, N3, O4)	02*, G(O6), G(N2)
Ratio of the base to backbone interaction	0.64	1.0	0.72	1.1
Main chain interaction	26%	–	–	26%
Side chain interaction	74%	–	–	74%
van der Waals interactions (W)				
Proportion of W bonds in the study	35%	92%	–	72%
Frequent amino acids	Arg, Lys, Ser, Asn	Arg, Tyr, Phe, Asn, Thr	–	Arg, Asn
Rare amino acids	Val, lle, Pro, Trp	Cys, His, Pro, Ala	–	–
Frequent nucleotide–amino acid pairs	A(Ser, Arg, Thr) C (Arg, Lys, Tyr) G (Arg, Asn, Lys) U (Arg, Lys, Asn)	–	–	–
Nucleotides in the order of preference	G, C, U, A	G, U, A, C	–	–
Frequent protein atoms	N, ND2, NE2, NH1, NZ, NH2, O, OG1 , OG, OH	–	–	
Frequent RNA atoms	A(O2*, O3*, O1P, O2P), C(O2*, O3*, O1P, O2P), G(O2*, O6, O3*, O1P), U(O2, O4, N3)	O2*	–	O2*
Ratio of the base to backbone interaction	0.28	1.41	–	1.0
Main chain interaction	22%	–	–	–
Side chain interaction	78%	–	–	–

An entry with the '–' symbol indicates the data are not available.

nucleotide are responsible for most H-bonding interactions in our study, whereas these atoms are not significant in Allers and Shamoo's analysis. On the other hand, the interacting frequency of O2* in our study is quite similar with that in the studies by Treger and Weshtof [8] and Nadassy et al. [14]. The atoms O2*, O1P, and O2P are responsible for 53% of H bonding interactions in our analysis (Supplementary Tables 2–5).

Similarly in van der Waals interactions, the atoms O2*, O1P and O2P are the most prominent in all nucleotides, constituting 49% of the total van der Waals interactions (Supplementary Tables 7–10). In addition to these atoms, each nucleotide has its own preferred atoms, A (O3*), C (O3*, O2), G (O3*, O6) and U (O3*, O2) in van der Waals interactions.

In Treger and Weshtof's analysis [8], the preferred residues in H bonding interactions are Arg, Asn, Ser and Lys. In our analysis, Arg, Lys, Asn and Ser are the most frequent residues in H bonding interactions, and Arg, Lys, Asn and Gln have the highest H bonding interaction propensities. The least preferred residues in Treger and Weshtof's analysis are Ala, He, Leu and Val, while Cys, Val, Trp and Met are the least preferred residues in our analysis. The most favored pairs are A (Ile, Pro, Ser), C (Leu), G (Asp, Gly) and U (Asn) in their study, while the favored pairs are A (Arg, Lys, Ser), C (Arg, Lys, Ser), G (Arg, Lys, Glu) and U (Arg, Lys, Asn) in ours. The atom O2* is involved in 21% of H bonds in Treger and Weshtof's study, which is similar to 22% in our study. In Treger and Weshtof's study amino acids interact with nucleotides through main chains (26%) and side chains (74%), whereas in our work the corresponding figures are 26% and 74%. In van der Waals contacts, interactions occur in main chains (22%) and side chains (78%) in our study. Both Treger and Weshtof's study and our study showed that phosphate is preferred to ribose and that backbone is preferred to base in H bonds and van der Waals contacts.

Polar amino acids are the most interacting of all the amino acids. They constitute 69% of H bonds and 65% of van der Waals contacts in our study, and charged amino acids constitute about 55% of H bonds and 51% of van der Waals contacts. Charged amino acids contribute to 40% of interactions in Treger and Weshtof's study. Positively charged amino acids are more interacting than negatively charged ones in both studies. The interactions of hydrophobic amino acids are quite frequent (22%) in their work, while they are negligible (5%) in our work. Aliphatic and aromatic amino acids also form a considerable number (13%) of interactions in our work, while their interactions are not noted in their work. The differences in interaction frequencies are due to different definitions of interaction. The sum of the percentages exceeds 100% in our work, since the classifications of amino acids are not mutually exclusive, i.e., the same amino acid can be assigned to more than one group.

In summary, we have described the development of a web-based program called PRI-Modeler, and an analysis of data on the structure of protein–RNA complexes using the program to identify interactions between proteins and RNAs. We have analyzed a most representative set of 45 protein–RNA complexes and discovered several interaction patterns. PRI-Modeler can extract information about the secondary structure and tertiary structure of RNA and identify the interaction patterns between proteins and RNAs. We believe that PRI-Modeler will be helpful in research aimed at identifying the binding sites in protein–RNA complexes.

Acknowledgments: We thank Daeho Lim for the earlier work on PRI-Modeler. This work was supported by INHA UNIVERSITY Research Grant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2007.03.085.

References

[1] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structures: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22, 2577–2637.
[2] 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.
[3] Tinoco Jr. (1993) in: The RNA World (Gesteland, R.F. and Atkins, J.F., Eds.), pp. 603–607, Cold Spring Harbor Laboratory Press.
[4] McDonald, I.K. and Thornton, J.M. (1994) Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238, 777–793.
[5] Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M. and Thornton, J.M. (2001) Protein- RNA interactions: a structural analysis. Nucleic Acids Res. 29, 943–954.
[6] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[7] Nagaswamy, U., Voss, N., Zhang, Z. and Fox, G.E. (2000) Database of non-canonical base pairs found in known RNA structures. Nucleic Acids Res. 28, 375–376.
[8] Treger, M. and Westhof, E. (2001) Statistical analysis of atomic contacts at RNA–protein interfaces. J. Mol. Recog. 14, 199–214.
[9] Moodie, S.L., Mitchell, J. and Thornton, J. (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. J. Mol. Biol. 263, 468–500.
[10] Hubbard, S.J. and Thornton, J.M. (1993). NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College, London.
[11] Akmaev, V.R., Kelley, S.T. and Stormo, G.D. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. Bioinformatics 16, 501–512.
[12] Shi, H. and Moore, P.B. (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited. RNA 6, 1091–1105.
[13] Per, J.K. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24, 946–950.
[14] Nadassy, K., Wodak, S.J. and Janin, J. (1999) Structural features of protein–nucleic acid recognition sites. Biochemistry 38, 1999–2017.
[15] Allers, J. and Shamoo, Y. (2001) Structure-based analysis of protein–RNA interactions using the Program Entangle. J. Mol. Biol. 311, 75–86.