Mol. Cells, Vol. 16, No. 2, pp. 161-167

Molecules and Cells

Discovering the Interaction Propensities of Amino Acids and Nucleotides from Protein-RNA Complexes

Euna Jeong, Hyunwoo Kim, Seong-Wook Lee[1], and Kyungsook Han*

School of Computer Science and Engineering, Inha University, Incheon 402-751, Korea;
[1] Department of Molecular Biology, Dankook University, Seoul 140-714, Korea.

(Received March 16, 2003; Accepted May 26, 2003)

With the availability of many genome sequences, the mining of biological data is attracting much attention, most of it limited to the sequences of macromolecules. Sequence data are easy to analyze as they can be treated as strings of characters, whereas the structure of a macromolecule is much more complex. We developed a set of algorithms to analyze the structures of protein-RNA complexes at the atomic level and used them to analyze protein-RNA interactions using structural data on 51 protein-RNA complexes. The analysis revealed, among other things, that: (1) polar and charged amino acids have a strong tendency to interact with nucleotides, (2) arginine and asparagine tend to hydrogen bond with uracil, and (3) histidine favors uracil in water-mediated bonding with RNA. We analyzed a large set of structural data of protein-RNA complexes involving water-mediated hydrogen bonds as well as direct hydrogen bonds. The interaction patterns discovered from the analysis provide useful information for predicting the structure of RNA that binds proteins, and of proteins that bind RNA.

Keywords: Hydrogen Bond; Interaction Propensity; Protein-RNA Interaction; Structural Data.

Introduction

As the DNA sequence and genes of the human genome become known, discovering useful patterns from the data has become an important challenge. Much of the data mining in bioinformatics is limited to DNA, RNA and protein sequences. These are easy to analyze and there are many well-developed algorithms to handle them, since they can be treated as strings of characters. The structure of a macromolecule is much more complex. There have been few attempts to mine the three-dimensional structures of macromolecules, because much less is known about these and because there are no readily usable models or algorithms. However, an increasing number of the three-dimensional structures are being solved. The Protein Data Bank (PDB) (Berman *et al.*, 2000), for example, has currently 19,000 entries on structure data. Three-dimensional structure data on nucleic acids and proteins contain the x, y, and z coordinate values of all the atoms, and information on embedded elements such as hydrogen bonds and secondary structures.

Protein-DNA complexes have been investigated for many years (Deng *et al.*, 1999; Luscombe *et al.*, 2001), but protein-RNA complexes have received much less attention despite their importance. In contrast to the regular helical structure of DNA, RNA molecules form complex secondary and tertiary structures with stems, loops, and pseudoknots that are often recognized by specific proteins. RNA structures contain hydrogen bonds and electrostatic and hydrophobic groups that can form specific contacts with small molecules. However their specific interactions with proteins are not well understood.

As an extension of our previous study of 29 protein-RNA complexes (Kim *et al.*, 2002), we attempted to extract common features of protein-RNA complexes at the level of residues and atoms from a more comprehensive data set. The primary focus of our work is to establish how proteins selectively bind specific sites on RNA molecules. Thus, we calculated the hydrogen bond propensities not only of the 20 amino acids in proteins but also the 4 nucleotides in RNAs. We attempted to address the following problems:

- Which protein residues bind preferentially to which nucleotides?
- What are the interface properties between protein and

* To whom correspondence should be addressed.
Tel: 82-32-860-7388; Fax: 82-32-863-4386
E-mail: khan@inha.ac.kr

RNA in direct hydrogen bonding and in water-mediated hydrogen bonding?
- Are main chain contacts observed with the same frequency as side chain contacts in proteins?
- Are backbone contacts observed with the same frequency as base contacts in RNAs?
- Which protein atoms and which RNA atoms tend to be present at protein-RNA binding sites, in terms of hydrogen bond acceptors and donors?

Identifying hydrogen bonds

In this section we describe the method used to identify hydrogen bonds that do or do not involve water. We then calculate interaction propensities using a modification of the propensity function of Moodie *et al.* (1996).

Dataset generation Protein-RNA complex structures were obtained from the PDB database (Berman *et al.*, 2000). Complexes solved by X-ray crystallography at a resolution better than 3.0 Å were selected. As of September 2002, there were 188 protein-RNA complexes in the PDB database, 139 of them at a resolution of 3.0 Å or better. We used PSI-BLAST (Altschul *et al.*, 1997) for similarity searches on the protein and RNA sequences in these 139 complexes to eliminate equivalent amino acids or nucleotides in homologous protein or RNA structures. Complexes that do not contain water molecules were eliminated from the dataset. 51 of the 139 protein-RNA complexes were left as representative, non-homologous interactions after running PSI-BLAST with an E value of 0.001 and identity value of 80% or below. Table 1 lists these complexes.

Since we included any protein or RNA sequence with an identity value of 80% or lower in the dataset some proteins and RNAs are present in more than one complex. Our experience is that a protein molecule (or RNA molecule) shows different behavior at the atomic level with different partners. In our dataset, 15 complexes with tRNAs contained similar RNAs (both in terms of the RNA sequence and RNA structure). Other complexes contain similar protein sequences and structures. Table 1 classifies the protein-RNA complexes, and complexes with similar partners (in terms of sequence and structure) are placed in parentheses.

Hydrogen bonds A hydrogen bond is formed by three atoms: one hydrogen atom and two electronegative atoms (often N or O). The hydrogen atom is covalently bound to one of the electronegative atoms, the hydrogen bond donor; the other electronegative atom is known as the hydrogen bond acceptor. The electronegative atoms may take up some of the electron density from the hydrogen atom, and, as a result, each electronegative atom carries a partial negative charge and the hydrogen atom a partial positive charge. Consequently, the hydrogen atom and the hydrogen bond acceptor attract one another. The strength of the hydrogen bond depends on the donor and acceptor as well as their environment, the bond energy usually ranging from 1 to 5 kcal/mol. This energy is smaller than covalent bond energy, but greater than thermal energy (0.6 kcal/mol at room temperature). Therefore, a hydrogen bond can provide a significant stabilizing force.

Table 1. Protein-RNA complexes in the data set. complexes sharing similar interacting partners are in parentheses.

Molecular	No.	PDB ID
tRNA	15	(1EFW, 1SER, 1C0A, 1QTQ) (1H4Q, 1H4S) (1FFY, 1GAX) 1B23, 1F7U, 1G59, 1IL2, 1QF6, 2FMT, 1K8W
mRNA	1	1B7F
Ribosome	6	(1HC8, 1MMS) (1DFU, 1FEU) 1DK1, 1I6U
Ribozyme	4	(1JBR, 1JBS) 1CX0, 1B2M
Virus	12	(2BBV, 1F8V) 1KNZ (1E7X, 1HE0, 1HE6, 1HDW, 1ZDH, 1ZDI 5MSF, 6MSF, 7MSF)
TRAP	3	(1C9S, 1GTF, 1GTN)
SRP	4	(1JID, 1L9A, 1LNG) 1HQ1
Others	6	(1FXL, 1G2E) 1DI2, 1EC6, 1KQ2, 1URN

The number of hydrogen bonds between amino acids and nucleotides in the protein-RNA complexes was calculated using CLEAN, a program for tidying Brookhaven files, and HBPLUS version 3.15 (McDonald and Thornton, 1994), a program to calculate the number of hydrogen bonds. Hydrogen bonds were identified by finding all proximal atom pairs that satisfied given geometric criteria between hydrogen bond donors (D) and acceptors (A). The positions of the hydrogen atoms (H) were inferred from the surrounding atoms, as hydrogen atoms are invisible in purely X-ray-derived structures. The criteria considered for forming 4 hydrogen bonds in this study were: contacts with a maximum D-A distance of 3.9 Å, maximum H-A distance of 2.5 Å, and minimum D-H-A and H-A-AA angels set to 90, where AA is an acceptor antecedent.

Water-mediated ligand interactions are essential in biological processes. The presence of water at the interface is revealed by methods such as nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography. The tendency of nucleotides that bind water strongly was also confirmed by our analysis (see Table 3 shown later). One nucleotide can bind one or more water molecule. A water-mediated bond means that one water molecule forms a hydrogen bond with an amino acid on one side and a nucleotide on the other side. Figure 1 presents examples of

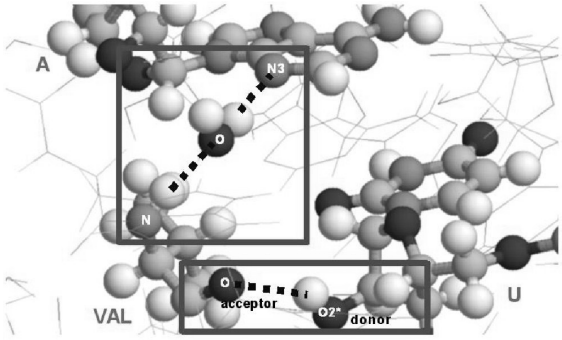

Fig. 1. Three-dimensional diagram of hydrogen-bonding atoms around an amino acid, and bases, by Rasmol. Dotted lines indicate hydrogen bonds. In the upper rectangle is a water-mediated bond between N (valine) and N3 (adenine). A direct bond between O (valine) and O2* (uracil) is shown in the lower rectangle.

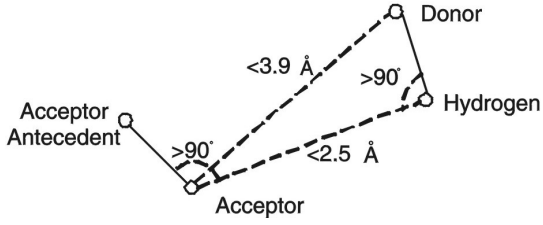

Fig. 2. Angles and distances used in the definition of the hydrogen bonds.

hydrogen bonds used in our study.

All protein-RNA bonds were extracted from the HBPLUS output files. There were 1568 hydrogen bonds and 1276 water-mediated hydrogen bonds in the dataset. In order to understand the properties of these bonds, we analyzed separately the results for two classes of protein-RNA complexes: (1) those with directed bonds and (2) those with water-mediated bonds.

Extraction processes for direct bonds and water-mediated bonds are shown in Fig. 3, and Algorithm 1 describes the pseudo code for extracting direct protein-RNA bonds and water-mediated bonds. The D-List contains direct bonds between proteins and RNA, and the WM-List water-mediated bonds.

Algorithm 1 to extract direct and water-mediated bonds
1: Given hydrogen bonds DATA *H*
2: **for all** bond *b* in *H* **do**
3: **if** *b* is bond between amino acid and nucleotide **then**
4: *D-List* ← *D-List* U *b*
5: **else if** b is bond between water and amino acid **then**
6: *W-A-List* ← *W-A-List* U *b*
7: **else if** *b* is bond between water and nucleotide **then**
8: *W-N-List* ← *W-N-List* U *b*
9: **end if**
10: **end for**

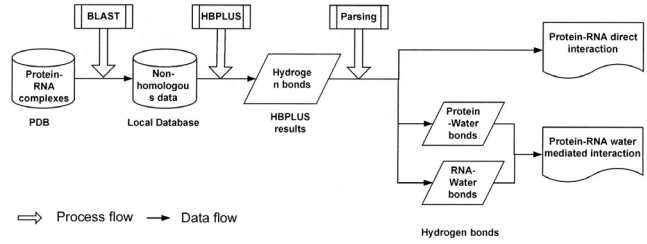

Fig. 3. The procedure for extracting direct and water-mediated protein-RNA interactions.

11:
12: **for all** water-amino acid bond *wab* in *W-A-List* **do**
13: let water molecule of *wab* be *wmol*
14: let amino acid of *wab* be *amol*
15: **for all** water-nucleotide bond *wnb* in *W-N-List* **do**
16: let nucleotide of *wnb* be *nmol*
17: **if** *wmol* is water molecule of *wnb* **then**
18: *WM-List* ← *WM-List* U (*amol*, *nmol*)
19: **end if**
20: **end for**
21: **end for**

Interaction propensity We define the interaction propensity (P) for each of the 20 common amino acids binding to each of the 4 nucleotides (adenine, guanine, cytosine, and uracil). Our propensity function is based on that of Moodie *et al*. (1996), modified to

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

where N_{ab} is the number of amino acids a in contact with nucleotide b, ΣN_{ij} is the total number of amino acids in contact with any nucleotide, N_a is the total number of amino acid a in the dataset, N_b is the total number of nucleotide b in the dataset, ΣN_i is the total number of amino acids in the dataset, and ΣN_j is the total number of nucleotides in the dataset. The numerator represents the ratio of hydrogen bonds between amino acid a and nucleotide b to total hydrogen bonds, while the denominator represents the ratio of frequencies of amino acid a and nucleotide b to those of all amino acids and nucleotides in the dataset. Hence, a propensity greater than 1 indicates that a given amino acid occurs more frequently in the protein-RNA interface with a given nucleotide than in the remainder of the protein surface, whereas a propensity less than 1 indicates that a given amino acid occurs less frequently in the interface with a given nucleotide.

The interaction propensity value in equation (1) is calculated as the proportion of a particular amino acid binding a particular nucleotide in the interface divided by the proportion of them in the data set. Therefore, the propen-

Table 2. Direct hydrogen bonds and water-mediated hydrogen bonds for each amino acid. Number in the left column are direct bonds, those in the right column water-mediated bonds.

Amino acids	Freq.	Nucleotides								Sub total		Total
		A		G		C		U				
ARG	1692	54	28	63	52	102	49	87	33	306	162	468
LYS	1716	100	21	50	40	58	41	49	15	257	117	374
SER	1598	48	59	51	41	36	34	29	36	164	170	334
THR	1582	76	88	44	37	9	33	22	12	151	170	321
ASN	1100	17	19	32	37	30	13	46	32	125	101	226
GLU	2072	12	12	96	26	21	26	7	17	136	81	217
ASP	1427	2	6	68	37	37	30	9	21	116	94	210
TYR	816	13	10	10	22	21	27	15	19	59	78	137
GLN	966	4	11	24	19	14	18	19	4	61	52	113
GLY	2058	6	10	18	18	13	9	3	18	40	55	95
HIS	631	8	18	15	13	4	5	9	8	36	44	80
LEU	2279	3	2	7	20	7	10	2	15	19	47	66
PHE	1145	0	6	31	5	0	3	0	2	31	16	47
ALA	2370	3	6	4	8	3	6	7	0	17	20	37
ILE	1585	6	1	2	3	2	10	0	3	10	17	27
VAL	2241	0	7	0	2	3	4	0	10	3	23	26
MET	514	3	2	4	9	1	1	1	3	9	15	24
PRO	1208	5	0	4	4	3	3	0	3	12	10	22
TRP	324	3	3	0	0	6	0	3	0	12	3	15
CYS	248	4	1	0	0	0	0	0	0	4	1	5
TOT	27572	367	310	523	393	370	322	308	251	1568	1276	2844

sity value can reveal the frequency of co-occurrences of amino acids and nucleotides in the protein-RNA complexes for every combination of amino acids and nucleotides.

Results

As described in Section 2, protein-RNA interactions (in terms of hydrogen bonds) were calculated for 51 complexes extracted from PDB using the program HBPLUS. In this section, we analyze 1568 direct bonds and 1276 water-mediated bonds in the dataset.

Characteristics of hydrogen bonds Table 2 shows the number of hydrogen bonds that do and do not involve water between the 20 amino acids and 4 nucleotides. The amino acids are shown in the left-hand column and the nucleotides along the top. Amino acids are ordered by the number of interactions (total number of bonds in the last column) that they make. The number of direct hydrogen bonds is shown to the left of the parenthesis and that of water-mediated hydrogen bonds is in the parentheses from the third column to the seventh column. The last column gives the total number of bonds. Polar and charged residues, such as arginine, lysine, glutamic acid, and histidine make the largest number of hydrogen bonds, while buried and hydrophobic residues, cysteine, methionine, isoleucine, and leucine, are used sparingly.

Frequencies and ratios of the four nucleotides in the dataset as well as in contact interfaces are shown in Table 3. The nucleotides are shown in the first column. The second column shows how many of each of the 4 nucleotides occurred in the dataset. The third column is divided into two sub-columns for the frequencies (D) and the ratios (D/F) of the number of direct bonds, and those (W and W/F) of water-mediated bonds are shown in the fourth column. The 4 nucleotides showed similar tendencis to form water-mediated hydrogen bonds.

Analysis of interaction propensities The computed propensities for each of the 20 amino acids to interact with each of 4 nucleotides are displayed in Table 4 and plotted in Fig. 4 for direct hydrogen bonds, and in Table 5 and Fig. 5 for water-mediated hydrogen bonds. In Tables 4 and 5, P_a in the last column is the propensity of each

Table 3. Distribution of direct hydrogen bonds and water-mediated hydrogen bonds for each nucleotide. *D*, *W* and *T* denote frequencies and *D/F*, *W/F* and *T/F* denote the ratios of each frequency to *F*.

Nucleotides	Freq.	Direct HB		Water-mediated HB		Total	
	(F)	(D)	(D/F)	(W)	(W/F)	(T)	(T/F)
A	611	367	0.60	310	0.51	677	1.11
G	956	523	0.55	393	0.41	916	0.96
C	80	370	0.46	322	0.40	692	0.87
U	524	308	0.59	251	0.48	559	1.07
TOT	2894	1568	0.54	1276	0.44	2844	0.98

Table 4. Interaction propensities of 20 amino acids and 4 nucleotides involving direct hydrogen bonds. P_a denotes the propensity of each amino acid for all 4 nucleotides and P_b in the bottom row the propensity of each nucleotide for all 20 amino acids.

Amino acids	Nucleotides				P_a
	A	G	C	U	
ARG	2.66	1.98	3.82	4.99	3.18
LYS	4.85	1.55	2.14	2.77	2.63
ASN	1.29	1.55	1.73	4.06	2
SER	2.5	1.7	1.43	1.76	1.8
THR	4	1.48	0.36	1.35	1.68
ASP	0.12	2.54	1.64	0.61	1.43
TYR	1.33	0.65	1.63	1.79	1.27
GLU	0.48	2.47	0.64	0.33	1.15
GLN	0.34	1.32	0.92	1.91	1.11
HIS	1.06	1.27	0.4	1.39	1
TRP	0.77	0	1.17	0.9	0.65
PHE	0	1.44	0	0	0.48
GLY	0.24	0.47	0.4	0.14	0.34
MET	0.49	0.41	0.12	0.19	0.31
CYS	1.34	0	0	0	0.28
PRO	0.34	0.18	0.16	0	0.17
LEU	0.11	0.16	0.19	0.09	0.15
ALA	0.11	0.09	0.08	0.29	0.13
ILE	0.32	0.07	0.08	0	0.11
VAL	0	0	0.08	0	0.02
P_b	1.11	1.01	0.85	1.08	

amino acid to bond with any nucleotide and P_b in the bottom row is that of each nucleotide with any amino acid.

For direct hydrogen bonds, arginine has the highest propensity (P_a = 3.18), followed by lysine, asparagine and serine; the corresponding P_a values are 2.63, 2.00, and 1.80. Arginine is also the most frequently observed residue in DNA-binding antibodies (Park *et al.*, 2001). Common features of the amino acids with high propensity are

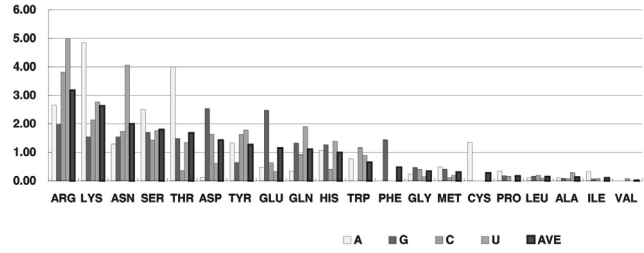

Fig. 4. Interaction propensities of amino acids and nucleotides in protein-RNA complexes with direct hydrogen bonds. The highest interaction propensity is observed for the arginine-uracil contact.

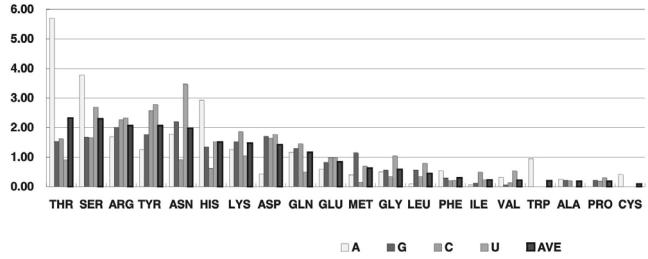

Fig. 5. Interaction propensities of amino acids and nucleotides in protein-RNA complexes with water-mediated hydrogen bonds. The highest interaction propensity is observed for threonine-adenine contacts.

that they are hydrophilic and contain highly electronegative atoms located in the outermost region of their side chains. In contrast, valine has the lowest propensity to interact with RNA (P_a = 0.02), followed by isoleucine (P_a = 0.11) and alanine (P_a = 0.13), and these amino acids are hydrophobic and do not have any highly electronegative atoms. It is noticeable that alanine and leucine, the most abundant amino acids in the dataset, form a very small number of hydrogen bonds (36 hydrogen bonds in total). Of the total number of 27572 amino acids, the frequencies of alanine and leucine are 2370 and 2279, respectively.

For water-mediated hydrogen bonds, the ranking of propensities differs slightly from that of direct hydrogen bonds. Threonine has the highest propensity (P_a = 2.32), followed by serine, arginine, and tyrosine, with *P* values of 2.30, 2.07, and 2.07, respectively. In contrast, cysteine has the lowest propensity to interact with RNA (P_a = 0.09), followed by proline (P_a = 0.18) and alanine (P_a = 0.18).

Comparing the results in Tables 4 and 5, lysine (P_a = 2.63 and 1.47) and arginine (P_a = 3.18 and 2.07) prefer direct bonds to water-mediated bonds, while tyrosine (P_a = 1.27 and 2.07) and threonine (P_a = 1.68 and 2.32) slightly favor water-mediated bonds. The rest of the amino acids and nucleotides show no preference. On the nucleotide side, guanine has the largest number of hydrogen bonds (916 hydrogen bonds) but this is mainly because guanine

Table 5. Interaction propensities of 20 amino acids and 4 nucleotides involving water-mediated hydrogen bonds. P_a in the last column denotes the propensity of each amino acid for all 4 nucleotides and P_b in the bottom row denotes the propensity of each nucleotide for all 20 amino acids.

Amino acids	Nucleotides				P_a
	A	G	C	U	
THR	5.69	1.53	1.62	0.91	2.32
SER	3.78	1.68	1.66	2.69	2.3
ARG	1.69	2.01	2.26	2.33	2.07
TYR	1.25	1.76	2.58	2.78	2.07
ASN	1.77	2.2	0.92	3.47	1.98
HIS	2.92	1.35	0.62	1.51	1.51
LYS	1.25	1.52	1.86	1.04	1.47
ASP	0.43	1.7	1.64	1.76	1.42
GLN	1.17	1.29	1.45	0.49	1.16
GLU	0.59	0.82	0.98	0.98	0.84
MET	0.4	1.15	0.15	0.7	0.63
GLY	0.5	0.57	0.34	1.04	0.58
LEU	0.09	0.57	0.34	0.79	0.45
PHE	0.54	0.29	0.2	0.21	0.3
ILE	0.06	0.12	0.49	0.23	0.23
VAL	0.32	0.06	0.14	0.53	0.22
TRP	0.95	0	0	0	0.2
ALA	0.26	0.22	0.2	0	0.18
PRO	0	0.22	0.19	0.3	0.18
CYS	0.41	0	0	0	0.09
P_b	1.15	0.93	0.91	1.09	

is the most abundant nucleotide (956 occurrences in the dataset). According to the ratio of total hydrogen bonds to the number of occurrences of each nucleotide, adenine is most favored, followed by uracil, guanine, and cytosine in direct bonds and water-mediated bonds. This relative order is also supported by the propensities computed for each nucleotide, as shown in the bottom rows of Table 4 and 5. Adenine has the highest propensity in both cases. However, guanine shows the largest difference between the propensities (P_b = 1.01 and 0.93) for direct and water-mediated bonds. This shows that guanine has little tendency to bind to proteins via water.

There are 7 clear preferences for particular pairings between amino acids and nucleotides in Figs. 4 and 5. Uracil strongly favors arginine (P = 4.99) and asparagine (P = 4.06) for direct hydrogen-bond interactions, and lysine and threonine favor adenine (P = 4.85 and 4.00, respectively). For water-mediated bonds, the threonine-adenine pair is the most common (P = 5.69). Threonine binds less to uracil, and tyrosine prefers cytosine, and asparagine favors uracil.

Amino acids are more diverse in their interaction propensities (average interaction propensity ranges from 0.02 to3.18) than nucleotides (range 0.85-1.11). Amino acids have a main chain in common, and nucleotides also have a backbone in common. In proteins, hydrogen bonds more often involve the side chain (71%) than the main chain (29%). In contrast, in RNA, hydrogen bonds in the backbone (51%) are slightly more frequent than in the RNA base (49%). Amino acids, in which side chain contacts are dominant, naturally have more diverse interaction propensities than nucleotides. One of the reasons that backbone contacts are more frequent than base contacts in nucleotides may be that the backbone has more atoms, and highly electronegative ones, than the base region.

Comparison with the hydrogen bond propensity of amino acids There are two important factors that determine the hydrogen bond propensity of an amino acid. One is the number of electronegative atoms and the other the accessibility of the atoms. The electronegative atoms have diverse accessibilities depending on their locations in amino acids. Electronegative atoms located in the outer area of an amino acid (for example, those in arginine and tyrosine) are easier to access and therefore form hydrogen bonds more frequently than those in the inner area (McDonald and Thornton, 1994).

The protein-RNA hydrogen bond propensities determined in our study showed a similar tendency. However, the hydrogen bonds between amino acids and nucleotides are also affected by the structures of the amino acids and nucleotides. For instance, amino acids with acidic side chains greatly prefer hydrogen bonding to guanine. The propensity of the Asp-G hydrogen bonds in direct hydrogen bonds is 2.54, much higher than the average propensity, 1.43, of aspartic acid (Table 4). Glutamic acid has a similar tendency: the propensity of the Glu-G hydrogen bond in direct hydrogen bonds is 2.47, higher than the average propensity, 1.15, of glutamic acid. This is because the N1 and N2 of guanine form stable hydrogen bonds to two oxygen atoms in aspartic acid and glutamic acid. This observation agrees with the recent analysis by Cheng *et al.* (2003) of hydrogen bond interactions between amino acid side-chains and nucleotide bases.

Analysis at the atomic level Essential atoms for hydrogen bonding exist both in the main chain and the side chain of proteins. However, side chain contacts are observed more frequently than main chain contact. The amino acids with main chain contacts have low interaction propensities and few electronegative atoms. These amino acids are also hydrophobic. In hydrophobic amino acids, the shorter the side chain, the higher its interaction propensity, because the side chain hinders the main chain from binding RNA.

Amino acids whose side chain contacts are generally much more dominant than main chain contacts have more diverse interaction propensities than nucleotides, in which backbone contacts are dominant. Highly electronegative atoms such as nitrogen and oxygen are very important in hydrogen bonding. Amino acids that contain polar atoms (e.g., arginine and lysine) have a strong tendency to form hydrogen bonds, while amino acids that contain an internal ring (e.g., tryptophan and histidine) have a weak tendency to form hydrogen bonds since the ring restricts the movement of the atoms. An exception to this is tyrosine; tyrosine indeed has a ring, but it has a moderate tendency to form hydrogen bonds because of the oxygen at the end of its side chain.

Conclusions

We have analyzed 2844 hydrogen bonds in the most representative set of 51 protein-RNA complexes, including both direct hydrogen bonds and water-mediated hydrogen bonds. The interaction propensity function developed for this analysis indicates the frequency of co-occurrences of amino acids and nucleotides in the protein-RNA complexes for every combination of amino acids and nucleotides. This interaction propensity function is more refined than others since our primary focus for the analysis is RNA as well as protein.

Amino acids, in which side chain contacts are much more dominant than main chain contacts, reveal more diverse interaction propensities than nucleotides, in which backbone contacts are dominant. On average, histidine has the highest propensity, followed by arginine, threonine, and lysine. These amino acids are hydrophilic and contain highly electronegative atoms in the outermost region of their side chains. In contrast, hydrophobic amino acids with no highly electronegative atoms, such as proline, isoleucine, and leucine, show a low propensity to interact with RNA. Amino acids other than tyrosine that contain an internal ring have a weak tendency to form hydrogen bonds. On the RNA side, uracil has the highest propensity, followed by adenine, guanine and cytosine. Our long-term goal is to predict the structure of RNA-binding protein, and we plan to extend this study to high-level structural analysis of protein-RNA complexes.

Acknowledgments This work was supported by the advanced backbone IT technology development program of the Ministry of Information and Communication of Korea under grant number 01-PJ11-PG9-01BT00B-0012.

References

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.

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.

Cheng, A. C., Chen, W. W., Fuhrmann, C. N., and Frankel, A. D. (2003) Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **327**, 781–796.

Deng, Y., Glimm, J., Wang, Y., Korobka, A., Eisenberg, M., and Grollman, A. P. (1999) Prediction of protein binding to DNA in the presence of water-mediated hydrogen bonds. *J. Mol. Model.* **5**, 125–133.

Kim, H., Jeong, E., and Han, K. (2002) Structural analysis of protein-RNA complexes. In *Proceedings of the Annual Meeting of Korean Society for Bioinformatics.* 28–38.

Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. (2001) Amino acid-base interactions: a three dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860–2874.

McDonald, I. K. and Thornton, J. M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.

Moodie, S. L., Mitchell, J., and Thornton, J. (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol.* **263**, 486–500.

Park, J., Kim, Y., Chung, H., Baek, K., and Jang, Y. (2001) Primary structures and chain dominance of anti-DNA antibodies. *Mol. Cells* **11**, 55–63.