An Algorithm for Predicting Binding Sites in Protein-Nucleic Acid Complexes

Namshik Han, Kyungsook Han*

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

***To whom correspondence should be addressed. E-mail: khan@inha.ac.kr**

Abstract

Determining the binding sites in protein-nucleic acid complexes is essential to the complete understanding of protein-nucleic acid interactions and to the development of new drugs. We have developed a set of algorithms for analyzing protein-nucleic acid interactions and for predicting potential binding sites in protein-nucleic acid complexes. The algorithms were used to analyze the hydrogen-bonding interactions in protein-RNA and protein-DNA complexes. The analysis was done both at the atomic and residue level, and discovered several interesting interaction patterns and differences between the two types of nucleic acids. The interaction patterns were used for predicting potential binding sites in new protein-RNA complexes.

Introduction

A variety of problems concerned with protein-DNA interactions have been investigated for many years, but protein-RNA interactions have been much less studied despite their importance. One reason for this is that only a small number of protein-RNA structures were known. As a result these structures were generally studied manually 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. Now that an increasing number of protein-RNA structures are known, there is a need to automatically analyze the interactions involved and to compare them with protein-DNA interactions.

In contrast to the regular helical structure of DNA, RNA molecules form complex secondary and tertiary structures consisting of elements such as stems, loops, and pseudoknots. Generally only specific proteins recognize a given configuration of such structural elements in three-dimensional space. RNA forms hydrogen bonds and electrostatic interactions, and possess hydrophobic groups; it can therefore make specific contacts with small molecules. However, the basis of its interaction with proteins is unclear. This paper presents a computational approach for analyzing the hydrogen-bonding interactions

between the amino acids of proteins and the nucleotides of nucleic acids and for predicting potential binding sites in protein-nucleic acid complexes.

Types of Hydrogen Bonding Interactions

Hydrogen bonds were classified into 3 types: (1) ***single interactions*** in which one hydrogen bond is found between an amino acid and a nucleotide, (2) ***bidentate interactions*** where an amino acid forms two or more hydrogen bonds with a nucleotide or base-paired nucleotides, and (3) ***complex interactions*** where an amino acid binds to more than one base step simultaneously (see Figure 1). Our definition of hydrogen bond types is slightly different from that of Luscombe *et al.* [1]. The latter only analyzed hydrogen bonds between amino acids and bases, whereas we also consider hydrogen bonds with the RNA backbone (sugar and phosphate). Therefore, our study can reveal differences in binding propensities between bases, sugar groups and phosphate groups.

Frameworks

Dataset

Protein-RNA complex structures were obtained from the PDB database [2]. Complexes solved by X-ray crystallography at a resolution ≤ 3.0Å were selected. As of September 2002, there were 188 protein-RNA complexes in PDB, and 139 of them were at a resolution ≤ 3.0 Å. We used PSI-BLAST [3] for similarity searches on each of the protein and RNA sequences in these 139 complexes in order to eliminate equivalent amino acids or nucleotides in homologous protein or RNA structures. 64 out of 139 protein-RNA complexes were left as the representative, non-homologous complexes after running the PSI-BLAST program with an E value of 0.001 and an identity value of 80% or below. We excluded 13 out of the 64 complexes that have no water molecules or are composed of artificial nucleotides. Therefore, the final data set was composed of 51 protein-RNA complexes. Table 1 shows the list of 51 protein-RNA complexes studied in our analysis. For the dataset of protein-DNA complexes, we used 129 protein-DNA complexes in the study of Luscombe [1].

Hydrogen Bonds

The number of hydrogen bonds between the amino acids and nucleotides in the protein-RNA complexes was calculated using CLEAN, a program for tidying Brookhaven files, and HBPLUS [4], a program to calculate the number of hydrogen bonds. The hydrogen bonds were identified by finding all proximal atom pairs between hydrogen bond donors (D) and acceptors (A) that satisfy the given geometric criteria. The positions of the hydrogen atoms (H) were theoretically inferred from the surrounding atoms, because hydrogen atoms are invisible in purely X-ray-derived structures. The criteria considered to form the hydrogen bonds for 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 angles set to 90°, where AA is an acceptor antecedent.

Table 1. The 51 protein-RNA complexes in the data set

PDB	Complex	Organism[a]	Res(A)	RNA residues	Protein residues[b]
1B23	EF-Tu-tRNA	*T.aquaticus*	2.60	74	405
1B2M	Ribonuclease T1	*A.oryzae*	2.00	2	104
1B7F	Sex-lethal	*D.melanogaster*	2.60	12	168
1C0A	Asp-tRNA synthetase	*E.coli*	2.40	77	585
1C9S	TRAP	*B.stearo.*	1.90	55	74
1CX0	U1A	*Hepatitis delta virus*	2.30	72	95
1DFU	L25	*E.coli*	1.80	19	94
1DI2	Protein A dsRBD	*X.laevis*	1.90	10	69
1DK1	S15-rRNA	*T.thermophilus*	2.80	57	86
1E7X	MS2 coat protein	*MS2*	2.38	19	129
1EC6	Nova KH	*H.sapiens*	2.40	20	87
1EFW	Asp-tRNA synthetase	*T.thermophilus*	3.00	73	580
1F7U	Arg-tRNA synthetase	*S.cerevisae*	2.20	76	607
1F8V	Mature capsid protein	*Pariacoto virus*	3.00	40	355
1FEU	L25	*T.thermophilus*	2.30	40	206
1FFY	Ile-tRNA synthetase	*S.aureus*	2.20	75	917
1FXL	HUD	*H.sapiens*	1.80	9	167
1G2E	HUD	*H.sapiens*	1.80	10	167
1G59	Glu-tRNA synthetase	*T.thermophilus*	2.40	75	468
1GAX	Val-tRNA synthetase	*T.thermophilus*	2.90	75	865
1GTF	TRAP	*B.stearo.*	1.75	55	74
1GTN	TRAP	*B.stearo.*	2.50	56	74
1H4Q	Pro-tRNA synthetase	*T.thermophilus*	3.00	77	477
1H4S	Pro-tRNA synthetase	*T.thermophilus*	2.85	77	477
1HC8	L11	*T.thermophilus*	2.80	58	76
1HDW	MS2 coat protein	*MS2*	2.60	19	129
1HE0	MS2 coat protein	*MS2*	2.68	19	129
1HE6	MS2 coat protein	*MS2*	2.65	19	129
1HQ1	SRP-4.5S RNA	*E.coli*	1.52	49	105
1I6U	S8-RRNA	*M.jannaschii*	2.60	37	130
1IL2	Asp-tRNA synthetase	*E.coli*	2.60	75	590
1JBR	Restrictocin-inhibitor	*A.restrictus*	2.15	62	149
1JBS	Restrictocin-inhibitor	*A.restrictus*	1.97	29	149
1JID	SRP	*H.sapiens*	1.80	29	128
1K8W	tRNA pseudouridine synthase	*E.coli*	1.85	22	327
1KNZ	NSP3 homodimer	*Rotavirus*	2.45	5	164
1KQ2	HFQ-RNA	*S.aureus*	2.71	7	77
1L9A	SRP	*M.jannaschii*	2.90	128	87
1LNG	SRP	*M.jannaschii*	2.30	97	87
1MMS	L11	*T.maritima*	2.57	58	140
1QF6	Thr-tRNA synthetase	*E.coli*	2.90	76	642
1QTQ	Glu-tRNA synthetase	*E.coli*	2.40	75	553
1SER	Ser-tRNA synthetase	*T.thermophilus*	2.90	94	421
1URN	U1A	*H.sapiens*	1.92	21	97
1ZDH	MS2 coat protein	*MS2*	2.70	19	129
1ZDI	MS2 coat protein	*MS2*	2.70	19	129
2BBV	Nucleocapsid	*Nodavirus*	2.80	10	363
2FMT	tRNA transformylase	*E.coli*	2.80	78	314
5MSF	MS2 coat protein	*MS2*	2.80	18	129
6MSF	MS2 coat protein	*MS2*	2.80	24	129
7MSF	MS2 coat protein	*MS2*	2.80	14	129

a *A.oryzae, Aspergillus oryzae; A.restrictus, Aspergillus restrictus; B.stearo., Bacillus stearothermophilus; D.melanogaster, Drosophila melanogaster; E.coli, Escherichia coli; H.sapiens, Homo sapaiens; M.jannaschii, Methanococcus jannaschii; S.aureus, Staphylococcus aureus; T.thermophilus, Thermus thermophilus; T.aquaticus, Thermus aquaticus; T.maritima, Thermotoga maritima; X.laevis, Xenopus laevis.*

b For NMR structures, one model was selected from the ensemble. NMR structures were evaluated on the basis of the r.m.s.d.

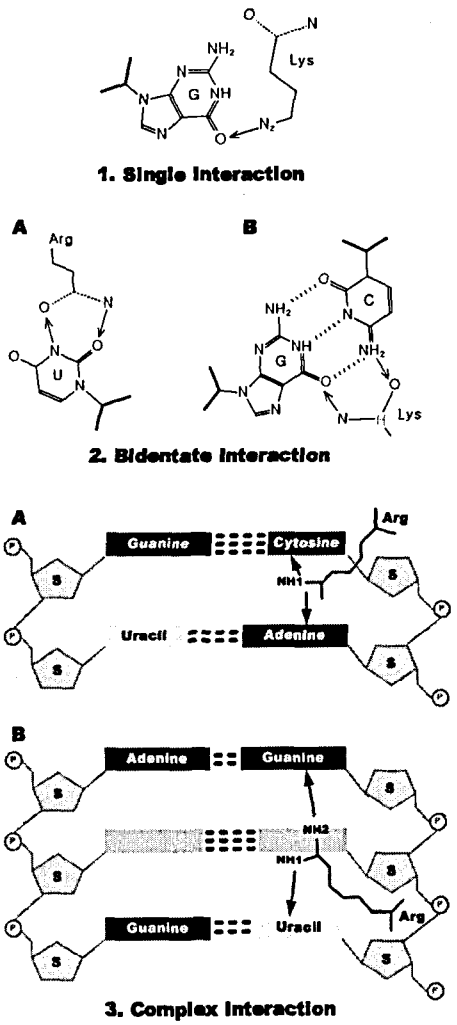

Figure 1. Schematic diagram of three interactions

All protein-RNA bonds were extracted from the HBPLUS output files. There were 1,568 hydrogen bonds in the dataset. We conducted separate experiments in order to compare the properties of single interactions, bidentate interactions and complex interactions, and the results were analyzed for the three types of hydrogen bonds: (1) single interactions, (2) bidentate interactions, and (3) complex interactions.

Algorithms

Analysis Algorithm

The analysis algorithm is composed of 4 phases (see Figure 2). In phase 1, the algorithm constructs the PRO-SEQ and NA-SEQ arrays to store the amino acid and nucleic acid sequences, respectively, and classifies hydrogen bonds into a P-N-List (list of hydrogen bonds between the protein and nucleic acid) and N-N-List (list of hydrogen bonds between nucleic acid). These arrays and lists are used to determine interaction types. The algorithm also analyzes whether a nucleotide is paired with other nucleotides. This process also uses the NA-SEQ and N-N-List. It is essential to discriminate whether binary or multi bond is single interaction or not. So, it is used to classify the interaction types. These processes are the basis of phases 2 – 4 of the algorithm.

In phase 2, the algorithm investigates the internal hydrogen bond relations of the nucleic acid and records the result of the investigation in a linked-list. It also investigates the hydrogen bonds between the protein and nucleic acid and records this result in a linked-list. These processes are not used to classify interaction types but represent important groundwork for identifying binding patterns as they represent the relation between pairs of residues in the form of linked-lists. These are then used in phase 4 to parse the classified interaction types.

In phase 3, the algorithm classifies the bonding type of each amino acid into unitary, double and multi-bond based on the number of hydrogen bonds between the amino acid and the nucleic acid. It inspects whether the amino acid forms two

or more hydrogen bonds with the base or base pair. This subroutine measures the distance between two nucleic acids by checking the chain and atom number of the each nucleic acid. The chain and atom number are included in PDB format. One nucleic acid of the two is searched in the P-N-List and the other is searched in the linked-list. So, if the two nucleic acids that are searched are the same, their distance is naturally zero. This is one of the most important processes because it can directly identify the double bond of the bidentate interaction. Since double bonds are abundant, it can eliminate many unnecessary operations. The algorithm classifies the protein-nucleic acid interaction types into three categories. These are single interactions, bidentate interactions and complex interactions. All unitary bonds belong to single interactions, and all double bonds belong to bidentate interactions. However, the classification of multi-bonds is not straightforward: if there are two or more hydrogen bonds with one base or base pair, they are classified as bidentate interactions. If there are two or more hydrogen bonds with more than one base step simultaneously, they are complex interactions. The rest are single interactions.

In phase 4, the algorithm parses the outcomes of phase 3 to determine binding patterns and numbers of hydrogen bonds involving each region of nucleotides and amino acids. The analysis is done both at the atomic and residue level, and the results help us identify how proteins recognize binding targets, which nucleotides are favored by which amino acids, and their binding sites.

Prediction Algorithm

The prediction algorithm is composed of two phases. In phase 1, it splits unknown protein structure into dices and examines all dices to sort potential binding sites with high probability. Splitting the protein structure requires the coordinate values of all atoms and the center position of every residue. Every PDB file of a structure has the starting coordinate value, which is outside the structure. The algorithm selects the closest residue from the starting coordinates of the structure. It then finds neighbor residues of the closest residue and the residues within a dice.

In phase 2, the algorithm constructs the structure-based residue lists that contain structural information for each dice. It then compares the lists to the nucleic acid sequence to predict potential binding sites using the interaction propensities and patterns. Finally, all potential binding sites are examined to predict the best binding site candidate. Figure 2 shows the sequence for classifying hydrogen bond types and for predicting binding sites.

Results

Differences between Protein-RNA and Protein-DNA Interactions

In protein-DNA complexes, almost equal numbers of hydrogen bonds were involved in single, bidentate and complex interactions [1]. However, in protein-RNA complexes, 60% of the hydrogen bonds were found in single interactions.

RNA and DNA were also different in their preference for backbone versus base contacts. 32% of the hydrogen bonds between protein and

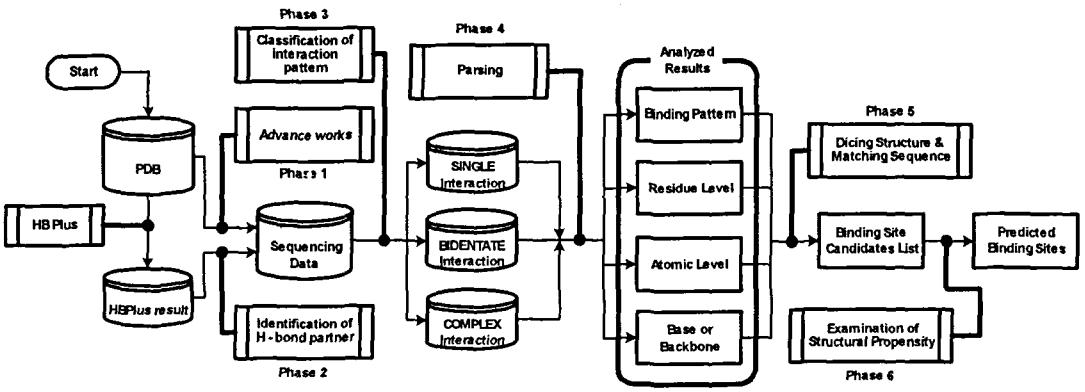

Figure 2. Sequence for classifying hydrogen bond types and for predicting binding sites

DNA involved base contacts, compared with 50% in protein-RNA. The fact that RNA bases bind to amino acids more frequently than do DNA bases can be explained by the structural difference between RNA and DNA. DNA is a double stranded molecule, and its bases are therefore already involved in hydrogen bonding. Hence, the base region is not as flexible as the backbone and is therefore less able to bind to amino acids. The bases in single-stranded regions of RNA, on the other hand, are quite flexible.

Amino Acids Favored by Nucleic Acids

GLU and ASP have acidic side chain groups, and more frequently hydrogen bind to RNA than to DNA. In protein-DNA complexes, these two amino acids are ranked 11th and 12th, respectively, among the 20 amino acids. But in protein-RNA complexes they are ranked 5 and 7th (Table 2). In particular, both GLU and ASP bind very frequently to guanine. Guanine binds to GLU eighty-nine times and to ASP sixty times in the protein-RNA complexes.

The opposite is the case with GLY and ALA, which bind to DNA more often than to RNA. They rank 10th and 14th, respectively in protein-RNA complexes, but 5th and 9th, in protein-DNA complexes (see Table 2). Both GLY and ALA have non-polar side chains, the smallest of the 20 amino acids, and residues with small side chains bind to double stranded DNA more easily than those with large side chains.

Table 2. Comparison of protein-DNA complexes with protein-RNA complexes in terms of the number of hydrogen bonds in amino acids

	DNA		RNA	
1	ARG	597	ARG	306
2	LYS	293	LYS	257
3	THR	292	SER	164
4	SER	207	THR	151
5	**GLY**	**168**	**GLU**	**136**
6	ASN	167	ASN	125
7	GLN	149	**ASP**	**116**
8	TYR	80	GLN	61
9	**ALA**	**71**	TYR	59
10	HIS	60	**GLY**	**40**
11	**GLU**	**53**	HIS	36
12	**ASP**	**19**	PHE	31
13	ILE	16	LEU	19
14	CYS	11	**ALA**	**17**
15	TRP	11	PRO	12
16	PHE	10	TRP	12
17	VAL	10	ILE	10
18	LEU	7	MET	9
19	PRO	2	CYS	4
20	VAL	3	MET	1

Table 3. Frequent binding patterns involved in bidentate interactions. The hydrogen bond donor (D) and hydrogen bond acceptor (A) are indicated in parentheses.

ARG						ASP		ASN		GLU		GLN						SER		THR	
NH1	NH2	NE	NH2	NH1	NH2	OD1	OD1	OD1	ND2	OE1	OE2	OE1	NE2	OE1	NE2	OE1	NE2	OG	OG	OG1	OG1
(D)	(D)	(D)	(D)	(D)	(D)	(A)	(A)	(A)	(D)	(A)	(A)	(A)	(D)	(A)	(D)	(A)	(D)	(A)	(D)	(D)	(A)
C		U		G		G		A		G		U		G		G		A		A	
N3	O2	O2	O2	O6	O6	N1	N2	N6	N1	N1	N2	N3	O4	N2	O2	N2	N3	N6	N1	N7	N6
(A)	(A)	(A)	(A)	(A)	(A)	(D)	(D)	(D)	(A)	(D)	(D)	(D)	(A)	(D)	(A)	(D)	(A)	(D)	(A)	(A)	(D)
11		2		1		12		1		37		2		1		1		1		18	

Interaction Propensities and Patterns of RNA

In bidentate interactions, GLU and ASP mainly bind to guanine whereas THR and LYS generally bind to adenine. This binding preference results in characteristic patterns of binding between the amino acid and nucleotide pairs. For example, the binding pattern shown in the GLU–G pair is most common (37 examples in the protein-RNA complexes, 89 hydrogen bonds). An exception is LYS: there are 69 hydrogen bonds between LYS and adenine bases, but there is no prominent binding pattern.

In protein-RNA complexes, the side chain of an amino acid binds to the only one base rather than base pairs or base steps. In contrast, there are many hydrogen bonds between a side chain and a base pair or base step in protein-DNA complexes [2]. This difference can again be explained by the structural difference between RNA and DNA. RNA structures contain less double-stranded regions than DNA, RNA has more unpaired bases, so the amino acids in protein-RNA complexes have a lower probability of binding to a base pair or base step than those in protein-DNA complexes.

Since RNA has more unpaired bases than DNA, it dose not provide the chances that amino acid can bind to the base pair or base step. Therefore, bidentate and complex interactions of protein-RNA complexes involve mainly the side chain and one base. Table 3 lists frequent binding patterns and their frequency in the dataset of 51 protein-RNA complexes.

Structural Propensities

Protein helices bind equally to nucleotide pairs and non-pairs in hydrogen-bonding interactions. In contrast, sheets prefer non-pairs to pairs, and turns prefer pairs to non-pairs. Non-pairs have been considered to have high interaction propensity in general, but our study found this is not the case since turns prefer pairs and helices show no preference. In protein-RNA complexes, this implies that sheets prefer to bind to RNA loops and turns prefer to bind to RNA stems [5].

Binding Sites

Figure 3 shows both the known binding sites of the NS5B part of Hepatitis C Virus (HCV) [6] and the predicted binding site by our algorithm. The predicted binding sites do not exactly correspond to the known binding sites, but are exclusively contained in the known binding sites. Table 4

Table 4. Sequences of the known and predicted binding sites. Residues in red color represent those common in the known and predicted binding sites. Residues in blue represent those in the known sites only.

Residue Number	213					218							225					230		
Residue Name	ASN	PRO	MSE	GLY	PHE	SER	TYR	ASP	THR	ARG	CYS	PHE	ASP	SER	THR	VAL	THR	GLU		
Residue Number	277					282									291					303
Residue Name	ARG	ARG	CYS	ARG	ALA	SER	GLY	VAL	LEU	THR	THR	SER	CYS	GLY	ASN	THR	LEU	.....	ALA	CYS
Residue Number	309							316			319							326		
Residue Name	GLN	ASP	CYS	THR	MSE	LEU	VAL	ASN	GLY	ASP	ASP	LEU	VAL	VAL	ILE	CYS	GLU	SER		
Residue Number	336				340												352			
Residue Name	LEU	ARG	VAL	PHE	THR	GLU	ALA	MSE	THR	ARG	TYR	SER	ALA	PRO	PRO	GLY	ASP			
Residue Number	362		364							371					376					
Residue Name	LEU	ILE	THR	SER	CYS	SER	SER	ASN	VAL	SER	VAL	ALA	HIS	ASP	ALA					

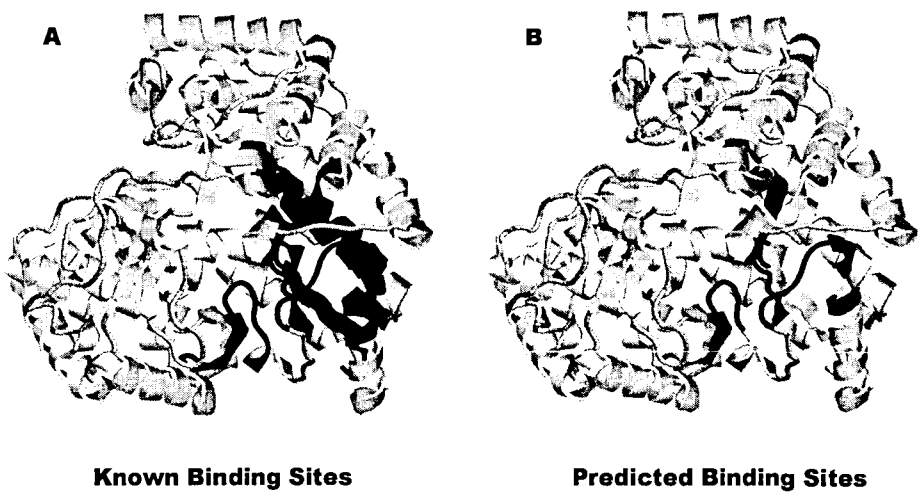

Figure 3. Binding sites in the NS5B part of Hepatitis C Virus

represents both predicted and known binding sites in sequence.

Discussion

We have developed a set of algorithms for analyzing hydrogen-bonding interactions between amino acids and nucleic acids and for predicting potential binding sites in protein-nucleic acid complexes. This paper presents the results of such an analysis and compares the characteristics of RNA and DNA binding to proteins.

DNA is a double-stranded molecule whereas RNA is usually single-stranded. This structural difference is the main cause of the difference in binding patterns of the two polynucleotides. The three hydrogen-bonding types were observed with equal frequency in DNA whereas single

interactions predominated in RNA. Backbone and base hydrogen bonds were observed with almost equal frequency in protein-RNA complexes (51% backbone hydrogen bonds and 49% base hydrogen bonds), but backbone hydrogen bonds were the majority in protein-DNA complexes (68% backbone hydrogen bonds and 32% base hydrogen bonds). DNA bonds involve GLY and ALA preferentially, whereas RNA usually does not bind to them but rather to GLU and ASP.

The protein-RNA complexes display specific binding patterns. In bidentate interactions, GLU and ASP overwhelmingly bind to guanine while THR and LYS generally bind to adenine. This binding preference results in favored binding patterns. For example, the binding pattern of the GLU–G pair is the most common (37 examples in the protein-RNA complexes with 89 hydrogen bonds). An exception is LYS; there are 69 hydrogen bonds between LYS and adenine bases, but no prominent binding pattern.

The binding patterns obtained from analyzing hydrogen-bonding interactions between amino acids and nucleotides were used to predict potential binding sites of Hepatitis C Virus. The binding sites predicted by our algorithm do not exactly correspond to the known binding sites of Hepatitis C Virus, but are exclusively included in the known binding sites. This indicates that prediction was performed in a conservative manner. However, a more rigorous study is required to improve the prediction results for various test cases.

Acknowledgements

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

References

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

[2] J. Westbrook, Z. Feng, L. Chen, H. Yang and H.M. Berman, The Protein Data Bank and structural genomics, *Nucleic Acids Research*, 31, 2003, 489-491

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. and Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25, 1997, 3389-3402

[4] I.K. McDonald and J.M. Thornton, Satisfying Hydrogen Bonding Potential in Proteins, J. *Mol.Biol.*, 238, 1994, 777-793

[5] H. Kim, E. Jeong, S.-W. Lee and K. Han, Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns, *FEBS Letters*, 552, 2003, 231-239

[6] S. Bressanelli, L. Tomei, A. Roussel, I. Incitti, R.L. Vitale, M. Mathieu, R. De Francesco and F.A. Rey, Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus, *Proc. Natl. Acad. Sci. USA*, 96, 1999, 13034-13039