Mining Molecular Structure Data for the Patterns of Interactions Between Protein and RNA

Kyungsook Han* and Chirag Nepal

School of Computer Science and Engineering, Inha University
Inchon, South Korea
khan@inha.ac.kr

Abstract. Mining useful information from a large amount of biological data is becoming important, but most data mining research in bioinformatics is limited to molecular sequence data. We have developed a set of algorithms for analyzing hydrogen bond and van der Waals interactions between protein and RNA. Analysis of the most representative set of protein-RNA complexes revealed several interesting observations: (1) in both hydrogen bond and van der Waals interactions, arginine has the highest interaction propensity, whereas cytosine has the lowest interaction propensity; (2) side chain contacts are more frequent than main chain contacts in amino acids, whereas backbone contacts are more frequent than base contacts in nucleotides; (3) amino acids, in which side chain contacts are dominant, reveal more diverse interaction propensities than nucleotides; and (4) valine rarely binds to any nucleotide. The interaction patterns found in this study should prove useful for determining binding sites in protein-RNA complexes.

1 Introduction

As the full genome sequences of several organisms are known, discovering useful information from the huge, noisy biological data has become an interesting problem. Many research of data mining in bioinformatics are limited to biological sequence data of macromolecules such as DNA, RNA and protein. Biological sequences are easy to analyze due to their sequential nature and have many well-developed algorithms since they can be treated as strings of characters. The structure of a molecule is much more complex than its sequence but plays an important role since it determines the biological function of the molecule. However, there have been few attempts to mine the structures of molecules, because much less is known about the structures than the sequences and there is no readily usable algorithm to handle structural data. However, an increasing number of the three-dimensional structures of molecules is known recently and continues to increase in public databases. The sequence of a molecule is a string of four bases and twenty amino acids, respectively, and is relatively simple to handle. The three-dimensional structure data of a molecule contains the x, y, and z coordinate values of all the atoms contained in the molecules and the information on embedded structure elements. Therefore, for the same molecule, its structure data is much larger and more complex than its sequence data.

* Corresponding author.

Over the past years a variety of problems concerned with protein-DNA complexes have been investigated (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 consisting of stems, loops, and pseudoknots. The structural elements arranged into three-dimensional space are often recognized by specific proteins. RNA structures display hydrogen bonding, electrostatic, and hydrophobic groups that can interact with small molecules to form specific contacts. However, it is not clear how proteins interact with RNA with specificity.

Since our previous study on hydrogen-bonding interactions between protein and RNA (Kim et al., 2003), the structures of many protein-RNA complexes have been determined. In this study we attempt a more rigorous study of the interactions by considering van der Waals interactions as well as hydrogen-bonding interactions in a new, extensive dataset of protein-RNA complexes. The primary focus of this work is to find how proteins selectively bind specific sites of RNA molecules in both types of interactions. The rest of this paper presents a set of algorithms for analyzing hydrogen-bonding and van der Waals interactions and the interaction patterns discovered from a new dataset of 45 protein-RNA complexes.

2 Identifying Hydrogen Bonds and van der Waals Contacts

2.1 Datasets of Protein-RNA Complexes

The protein-RNA complex structures were obtained from the PDB database (Berman et al. 2000). The complexes that were determined by X-ray crystallography with a resolution of 3.0 Å or better were selected. As of January 2006, there were 312 protein-RNA complexes in PDB and the number of complexes with a resolution of 3.0 Å or better was 207. We used PSI-BLAST (Altschul et al. 1997) for similarity search on each of the protein and RNA sequences in these 207 protein-RNA complexes, in order to eliminate the equivalent amino acids or nucleotides in the homologous protein or RNA structures. 45 out of 207 protein-RNA complexes were left as the representative, nonhomologous complexes after running the PSI-BLAST program with an E value of 0.001 and an identity value of 90% or below. The final dataset of 45 protein-RNA complexes was used in our study (Table 1).

Table 1. The list of 45 protein-RNA complexes in the data set

PDB ID									
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					

2.2 Hydrogen Bonds and van der Waals Contacts

The number of hydrogen bonds (H-bonds) between amino acids and nucleotides in protein-RNA complexes was calculated using CLEAN and HBPLUS version 3.15

(McDonald and Thornton 1994). The positions of hydrogen atoms (H) were inferred from the surrounding atoms, as 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 maximum D-A distance of 2.7 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. All protein-RNA bonds were extracted from the HBPLUS output files.

The criteria considered for the van der Waals interactions are: contacts with 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°. The interactions that are considered as H-bonds are not included in van der Waals interactions. The definitions of H-bonds and van der Waals interactions are slightly different from those used by Jones et al. (2001).

2.3 Binding 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 propensity, even though it may be involved in many H-bonds or may occur very frequently. Therefore, we employ a function for determining propensities that considers H-bond numbers, number of amino acids on the surface, nucleotide numbers and explosion values. The propensity function is based on that reported by Moodie et al. (1996), but we modified their function to determine the interaction propensities of amino acids and nucleotides pair 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 (Hubbard and Thornton 1993).

The interaction propensity P_{ab} between amino acid a and nucleotide b is defined by equation (1), where N_{ab} is the number of residues of amino acid a hydrogen bonding to nucleotide b , $\sum N_{ij}$ is the total number of amino acids hydrogen bonding to any nucleotide, N_a is the number of residues of amino acid a , $\sum N_i$ is the total number of amino acid residues, N_b is the number of nucleotide 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 all amino acids binding to any nucleotide on the surface. The term $N_a/\sum N_i$ of the denominator represents the ratio of the frequency of amino acid a to that of all amino acids on the surface, and the second term $N_b/\sum N_j$ represents the ratio of the frequency of nucleotide b to that of all nucleotides on the surface.

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

The interaction propensity is computed 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 represents the frequency of co-occurrence of amino acids and nucleotides in protein-RNA complexes, for every pair of amino acids and nucleotides. A propensity greater than 1 indicates that the given amino acid

occurs more frequently in combination with a given nucleotide on the surface, whereas a propensity less than 1 indicates that the amino acid occurs less frequently.

2.4 Structure Elements of Protein and RNA

An RNA nucleotide consists of sugar, phosphoric acid, and base. A base consists of a fixed number of atoms 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 hydrogen bonds between N1 of adenine (A) and N3 of uracil (U), and between N6 of A and O4 of U. Thus, we can establish definite rules for hydrogen bonding between fixed atoms, which we call base pair rules. 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 (Tinoco 1993) comprising of both canonical and non-canonical base pairs. Base pair rules are used for extracting base pairs and secondary and tertiary structural elements of RNA.

In this study, the secondary structure elements of protein are categorized into four types: helix (α -helix, 3/10 helix and π -helix), sheet (β -ladder and β -bridge), turn (hydrogen-bonded turn), and others (bend and other structures). These secondary structure elements are assigned using the DSSP program (Kabsch and Sander 1983). The secondary structure elements of RNA are categorized into two types: paired and unpaired. If at least one H-bond exists between the base atoms of two nucleotides, these nucleotides are considered to be paired. If not, they are considered to be unpaired. If an H-bond exists between other parts of two nucleotides than the base part, the two nucleotides are considered to be unpaired.

2.5 Algorithms for Extracting Secondary and Tertiary Structures of RNA

The first part of the algorithm extracts the secondary and tertiary structures of the RNA. The second part generates a visual representation of the RNA structure by integrating this information with knowledge of the coordinates of the nucleotides.

1. From the PDB files extract H-bonds that link the base of one nucleotide and that of another, and classify the H-bonds into the 28 types of base pairs using the base pair rules; they are then recorded in the Base Pair List.
2. Extract the RNA sequence data from the PDB and records it in RNA-SEQ. This contains information on all nucleotides in the RNA, which is used again in Step 4.
3. Derive the three-dimensional coordinates of the nucleotides. We define the average coordinate value of all the atoms that make up a nucleotide as the three-dimensional coordinate value of the nucleotide.
4. Integrate the sequence data in RNA-SEQ with the base pair data in the Base Pair List. All the nucleotides in RNA-SEQ are matched to the nucleotides in the Base Pair List to determine the bonding relationships between the nucleotides.
5. Integrate the three dimensional coordinates of the nucleotides obtained in Step 3 with the RNA structure data obtained in Step 4 and visualize the RNA structure.

2.6 Algorithms for Analyzing Protein-RNA Complexes

This algorithm analyzes the H-bonds and van der Waals contacts and generates the patterns of binding between protein and RNA.

1. Extract hydrogen bonds from the structure data of the protein-RNA complexes and assign secondary structure elements to each atom of the proteins.
2. Extract only hydrogen bonds that link RNA bases to get base pairs. Information on the base pairs obtained in this step, together with the information about the secondary structure elements of the protein obtained in Step 1, is used to identify the interaction pattern between protein and RNA at the secondary level.
3. Classify the amino acids into main chain and side chain. This can provide information on which type tends to form hydrogen bonds with nucleotides.
4. Show the analysis results between protein and RNA at various structure levels.

3 Results and Discussion

The algorithms were implemented in a web-based program called PRI-Modeler (<http://wilab.inha.ac.kr/primodeler/>) using Visual C#. PRI-Modeler is executable within a web browser on any Windows systems. Given one or more PDB files as input, PRI-Modeler produces several files containing H-bonding or van der Waals interactions between amino acids and nucleotides, secondary and tertiary structures of RNA, and interaction patterns of protein-RNA complexes.

Fig. 1. Analysis of the interactions between protein and RNA using PRI-Modeler

Fig. 1 shows an exemplary user interface of PRI-Modeler. It consists of twenty menu buttons and one text panel. When the user specifies the types of interactions to analyze using the menu buttons (such as H-bonding interactions, van der Waals interactions, hydrophobic interactions, aliphatic interactions, aromatic interactions, positively and negatively charged interactions and polar interactions), PRI-Modeler analyzes the specified interactions and displays the analysis results in the text panel. The rest of this section presents some analysis results of 45 protein-RNA complexes.

Table 2 shows the hydrogen bonds and van der Waals interactions between amino acids and nucleotides, and Fig. 2 shows their interaction propensities. There are a total

318 H-bonds and 1056 van der Waals contacts in the data set. Interestingly, both in the H-bonding and van der Waals interactions, arginine (Arg) showed the highest interaction propensity, whereas cytosine (C) showed the lowest interaction propensity (Fig. 2). Amino acids show more diverse nature (H-bonding interaction propensity in the range [0-2.44], and van der Waals interaction propensity in the range [0.04-3.16]) than nucleotides ([0.74-1.12] and [0.67-1.03], respectively). This indicates that amino acids are more distinguishable than nucleotides by their interaction propensities. Amino acids have a main chain in common, and nucleotides have a backbone in common. In protein, both interactions are more frequently observed in the side chain (average 73%) than in the main chain (average 27%). In contrast, the backbone part of RNA has more interactions (average 62% in both interactions) than in the base part (average 38%). Amino acids, in which side chain contacts are dominant, naturally reveal more diverse interaction propensities than nucleotides. One of the reasons that backbone contacts are more frequent than base contacts in nucleotides is that the backbone part has more atoms and highly electronegative atoms than the base part, which makes the backbone more favorable to form hydrogen bonds than the base.

Many amino acids show a strong preference for specific nucleotides with which they form stable H-bonds. For example, amino acids that contain polar atoms (e.g., arginine and lysine) have a strong tendency to form H-bonds. In H-bonding interactions, Arg-U, Gln-C, Arg-G, and Ser-A are the binding pairs with high propensity. In van der Waals interactions, Arg-U, Arg-G, and Asn-U are the binding pairs with high propensity. Valine rarely binds to any nucleotide.

Table 2. Hydrogen bonds and van der Waals interactions between amino acids and nucleotides

Amino acids	Frequency	hydrogen bonds					van der Waals interactions				
		A	C	G	U	Total	A	C	G	U	Total
Glu	3390	4	4	19	0	27	10	17	34	6	67
Lys	2866	9	8	8	16	41	39	55	50	26	170
Arg	2728	10	9	24	20	63	47	79	95	64	285
Leu	2270	0	0	0	1	1	5	8	0	1	14
Gly	2096	2	7	2	3	14	7	9	19	4	39
Asp	1960	8	1	10	8	27	1	14	26	8	49
Thr	1838	4	7	7	12	30	17	13	25	17	72
Ala	1764	0	1	0	2	3	2	8	6	8	24
Val	1688	1	0	0	0	1	0	3	0	0	3
Ser	1640	10	12	7	5	34	22	23	18	8	71
Pro	1628	3	0	0	0	3	1	4	2	0	7
Asn	1416	6	4	5	3	18	14	7	24	32	77
Gln	1290	2	13	3	4	22	11	12	28	19	70
Ile	1206	1	0	0	1	2	10	2	0	4	16
Tyr	1122	4	7	4	8	23	5	20	0	12	37
Phe	1032	0	0	0	0	0	1	0	12	3	16
His	870	0	5	4	0	9	6	8	5	2	21
Met	474	0	0	0	0	0	3	1	4	0	8
Trp	382	0	0	0	0	0	0	4	1	1	6
Cys	172	0	0	0	0	0	2	0	2	0	4
Total	31832	64	78	93	83	318	203	287	351	215	1056

Fig. 2. Average interaction propensities of amino acids (top) and nucleotides (bottom) for the hydrogen bonding and van der Waals interactions

Fig. 3. RNA structure extracted from a protein-RNA complex (PDB ID: 1EUY). H-bonds are shown in red dotted lines (left) and van der Waals contacts are shown green dotted lines (right).

Fig. 3 shows the RNA structures extracted from a protein-RNA complex (PDB ID: 1EUY). They are three-dimensional (3D) drawings, and can be rotated as well as zoomed in/out to get a clear perspective view. With programs like Rasmol (<http://www.umass.edu/microbio/rasmol/>) and MOLSCRIPT (Per 1991), one cannot easily obtain information about each nucleotide and the binding relations between nucleotides, because the programs represent the structures at the atomic level. On the other hand, PRI-Modeler shows not only the 3D configuration of RNA but also the bonding relations and types of base pairs between nucleotides.

4 Conclusion

We have analyzed 318 hydrogen bonds and 1056 van der Waals contacts in the most representative set of 45 protein-RNA complexes. The interaction propensity function we 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. Since our analysis is restricted to protein-RNA complexes that were crystallized and examined by X-ray crystallography, the analysis results may give rise to a high proportion of strong hydrogen bonding interactions. Our long-term goal of this study is to predict the structure of RNA binding protein, and we plan to extend this study to predict binding sites of virus proteins.

Acknowledgment. This work was supported by Inha University Research Grant.

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** (1997) 3389-3402
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. *Nucleic Acids Res.* **28** (2000) 235-242
3. Deng, Y., Glimm, J., Wang, Y., Korobka, A., Eisenberg, M., Grollman, A.P.: Prediction of Protein Binding to DNA in the Presence of Water-Mediated Hydrogen Bonds. *Journal of Molecular Modeling* **5** (1999) 125-133
4. Hubbard, S.J., Thornton, J.M.: NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, (1993) University College, London
5. Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M., Thornton, J.M.: Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.* **29** (2001) 943-954
6. Kabsch, W., Sander, C.: Dictionary of protein secondary structures: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22** (1983) 2577-2637
7. Kim, H., Jeong, E., Lee, S.-W., Han, K.: Computational Analysis of Hydrogen Bonds in Protein-RNA Complexes for Interaction Patterns. *FEBS Letters* **552** (2003) 231-239
8. Luscombe, N.M., Laskowski, R.A., Thornton, J.M.: Amino acid-base interactions: a three dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29** (2001) 2860-2874
9. McDonald, I.K., Thornton, J.M.: Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238** (1994) 777-793
10. Moodie, S.L., Mitchell, J., Thornton, J.: Protein recognition of adenylate: An example of a fuzzy recognition template. *J. Mol. Biol.* **263** (1996) 468-500
11. Per, J.K.: MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24** (1991) 946-950
12. Tinoco, Jr.: The RNA World. In: Gesteland, R.F., Atkins, J.F. (eds), Cold Spring Harbor Laboratory Press. (1993) 603-607