Mining Structure Elements from RNA Structure Data, and Visualizing Structure Elements

Daeho Lim, 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

Most currently known molecular structures were determined by X-ray crystallography or Nuclear Magnetic Resonance (NMR). These methods generate a large amount of structure data, even for small molecules, and consist mainly of three-dimensional atomic coordinates. These are useful for analyzing molecular structure, but structure elements at higher level are also needed for a complete understanding of structure, and especially for structure prediction. 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 small amount of structure data so far available, and extracting the structural elements of RNA requires substantial manual work. Since the number of three-dimensional RNA structures is increasing, a more systematic and automated method is needed. We have developed a set of algorithms for recognizing secondary and tertiary structural elements in RNA molecules and in the protein-RNA structures in protein data banks (PDB). The present work represents the first attempt at extracting RNA structure elements from atomic coordinates in structure databases. The regularities in the structure elements revealed by the algorithms should provide useful information for predicting the structure of RNA molecules bound to proteins.

Introduction

Mining biological data in databases has become the focus of increasing interest over the past several years. However most data mining in bioinformatics is limited to sequence data. The structure of a molecule is much more complex, but it is important as it determines the biological function of the molecule. It is therefore not enough just to analyze sequence data if one wishes to understand the structure of a molecule more completely. We have developed a set of algorithms that recognize secondary and tertiary RNA structure elements from the three-dimensional atomic coordinates of protein-RNA complexes. The algorithms are also able to use the information obtained to represent the structures visually.

Background

Base-Pair: An RNA nucleotide consists of a molecule of sugar, a molecule of phosphoric acid, and a molecule called a base. A base pair is formed when one base is paired with another base by hydrogen 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 [1] comprising both canonical and non-canonical base pairs. Figure 1 shows four base pairs.

Base-Pair Rules: A base consists of a fixed number of atoms (Fig 1). These fixed numbers provide important clues for extracting base pair data and classifying the data into types of base pairs. Base pairs are formed by hydrogen bonding between atoms of base. For example, the Watson-Crick 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 define the hydrogen bonds that generate base pairs and classify the base pairs. In this study we define base pair rules to classify base pairs, and divide them into 28 types [1] by means of these base-pair rules.

Methods

Our algorithm is divided into two parts. The first part derives information about secondary and tertiary structure elements of RNA by analyzing data in a PDB file [4]. We use HB-plus [5] to obtain data on all the hydrogen bonds that are present from the PDB file, and this data is used to generate base pair data. We can then obtain insight into the secondary or tertiary structure elements of the RNA by analyzing this data and integrating it with sequence data. The function of the second part is to derive a visual representation

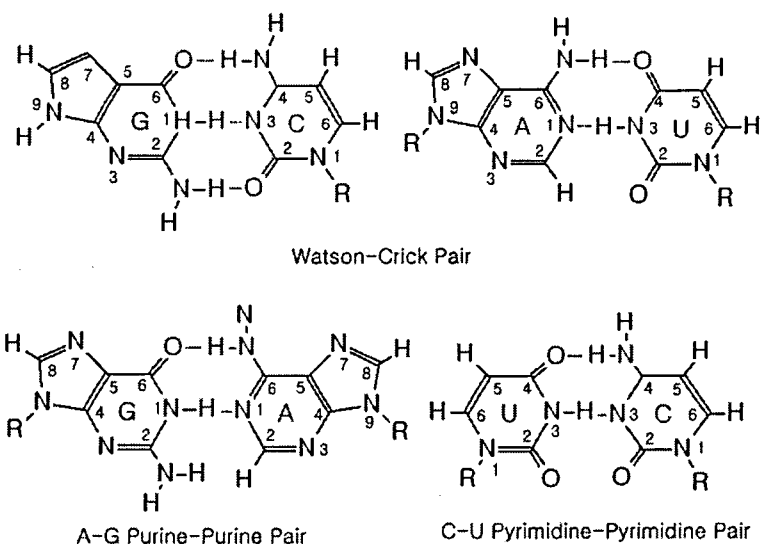

Fig 1. Watson-Crick G-C and A-U pairs, together with an A-G purine-purine pair, and a C-U pyrimidine-pyrimidine pair

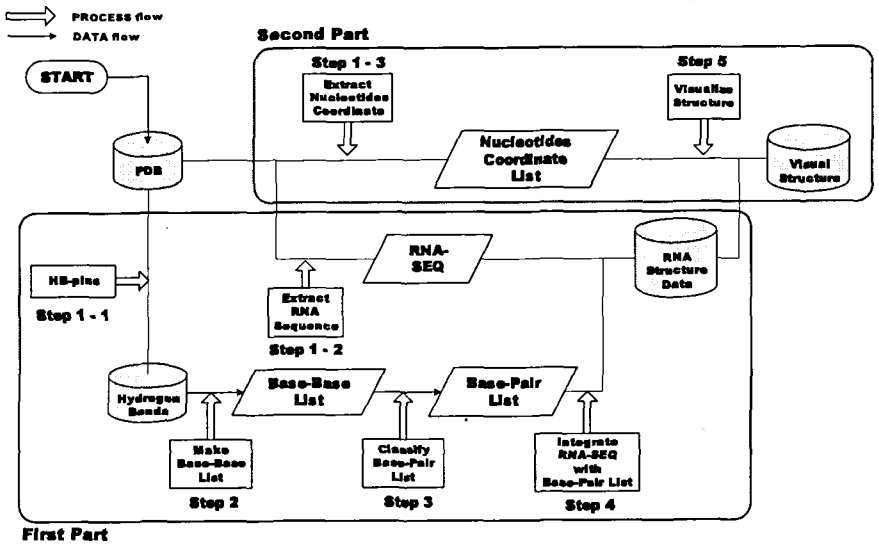

Fig 2. Framework for extracting information about secondary or tertiary structure elements of RNA from PDB and for visualizing the RNA structure

of the structure of the RNA by integrating the information about structure elements obtained in the first part with knowledge of the coordinates of the nucleotides. Fig 2 shows the framework of our approach.

Algorithm

The algorithm consists of the five steps shown in Fig 2.

Step 1: This step extracts data from a PDB file [4]. We obtain data on all the hydrogen bonds in the molecule in Step 1-1 by using HB-plus [5], and this data is recorded in Hydrogen Bonds. In Step 1-2 we extract the RNA sequence data by analyzing the PDB file, and record it in RNA-SEQ. Step 1-3 is a procedure for obtaining the 3D coordinates of the nucleotides. We define the average coordinate values of all the atoms by organizing each nucleotide by means of its 3D coordinates. The PDB file contains all the 3D coordinate values of the atoms.

Step 2: Step 2 is a process for extracting only those hydrogen bonds that bond one base to another from the hydrogen bond data obtained in Step 1-1. These hydrogen bonds are then recorded in the Base-Base List.

Step 3: Since not all the hydrogen bonds in the Base-Base List are involved in base pairing this step extracts those hydrogen bonds that are so involved, and classifies them into the 28 types by means of the Base pair rules. These hydrogen bonds are then recorded separately in the Base-Pair List.

Step 4: This step integrates the sequence data in RNA-SEQ with the base pairs data in the Base-

Pair List; the algorithm matches all the nucleotides in RNA-SEQ to the nucleotides in the Base-Pair List to determine the hydrogen bonding relationships of each nucleotide. The final output of Step 4 is information about the secondary and tertiary structure elements of the RNA.

Step 5: Step 5 represents the structure of the RNA visually by combining the information about structure elements obtained in step 4 with the coordinate values of the nucleotides obtained in steps 1-3.

Table 1. Information on the tertiary structure elements of MMTV from a PDB file (PDB identifier: 1RNK)

Nucleotide		Nucleotide to be paired		
Number	Symbol	Number	Symbol	Base-pair
0001	G	0019	C	G-C Watson Crick
0002	G	0018	C	G-C Watson Crick
0003	C	0017	G	G-C Watson Crick
0004	G	0016	C	G-C Watson Crick
0005	C	0015	G	G-C Watson Crick
0006	A			
0007	G			
0008	U	0033	A	A-U Watson Crick
0009	G	0032	C	G-C Watson Crick
0010	G	0031	C	G-C Watson Crick
0011	G	0030	C	G-C Watson Crick
0012	C	0029	G	G-C Watson Crick
0013	U	0028	G	G-U Wobble Pair
0014	A			
0015	G	0005	C	G-C Watson Crick
0016	C	0004	G	G-C Watson Crick
0017	G	0003	C	G-C Watson Crick
0018	C	0002	G	G-C Watson Crick
0019	C	0001	G	G-C Watson Crick
0020	A			
0021	C			
0022	U			
0023	C			
0024	A			
0025	A			
0026	A			
0027	A			
0028	G	0013	U	G-U Wobble Pair
0029	G	0012	C	G-C Watson Crick
0030	C	0011	G	G-C Watson Crick
0031	C	0010	G	G-C Watson Crick
0032	C	0009	G	G-C Watson Crick
0033	A	0008	U	A-U Watson Crick
0034	U			

Results Discussion

Table 1 presents the information on the tertiary structure elements of mouse mammary tumor virus (MMTV) extracted by our algorithm from a PDB file (PDB identifier: 1RNK). Nucleotides in columns 3 and 4 base pair with nucleotides in columns 1 and 2. Column 5 gives the type of base pair involved. If a nucleotide does not pair with another nucleotide, columns 3, 4 and 5 are left blank. Fig 3 (1) displays the structure of MMTV derived by our algorithm from the information in Table 1. Nodes indicate nucleotides of RNA and the blue lines indicate that nucleotides are connected in the RNA backbone. In addition the red dotted lines indicate that two bases are hydrogen bonded. Fig 3 (1) is a pseudoknot, i.e. a tertiary structure element formed when the bases in a single-stranded loop hydrogen with bases outside the loop. This pseudoknot conforms to the known structure of MMTV.

Our algorithm can also extract structure elements involving separate RNA strands by using hydrogen bonds data to obtain base pair data. Fig 3 (2) shows two RNA chains (chains M and N), extracted from a protein-RNA complex (PDB identifier: 1DFU). It can also identify base-triplets. A base-triplet is a tertiary RNA interaction in which a base pair interacts with a third base [8]. Fig 5 shows the secondary structure of a tRNA (PDB identifier: 1EHZ), obtained by our algorithm. Each of the base-triplets labeled 1 and 2 in Fig. 4 corresponds to a base-triplet (see Fig 5 for the structure of the base-triplets at the atomic level). A base-triplet has both secondary and tertiary interactions. For example, in the C-G-G base-triplet of Figs 4 and 5, the G-C pair is a secondary interaction in the form of a Watson-Crick pair, and the G-G pair is a tertiary interaction in the form of a purine-purine pair.

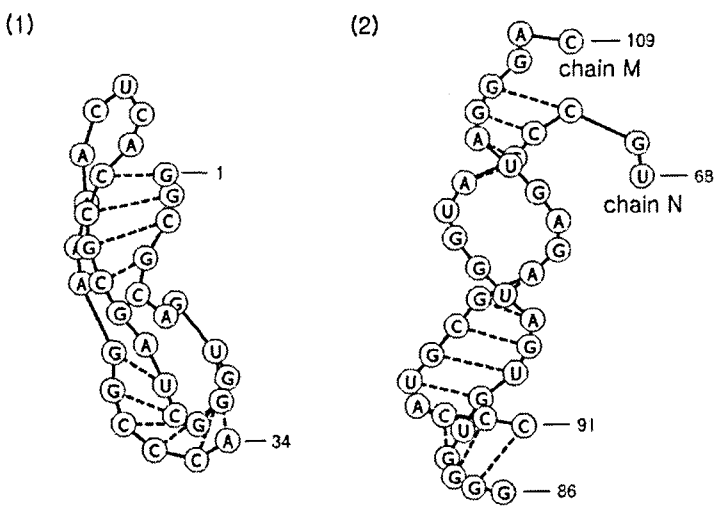

Fig 3. (1) Pseudoknot structure of MMTV, formed by the base pairs listed in Table 1. (2) Structure consisting of two RNA chains (M and N), extracted from a protein-RNA complex (PDB identifier: 1DFU). These structures were represented visually by our algorithm

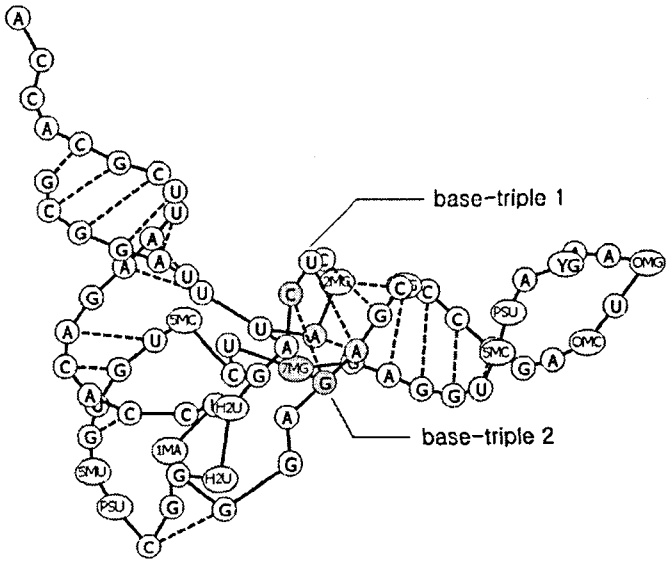

Fig 4. Structure of tRNA in a protein-RNA complex (PDB identifier: 1EHZ). This figure shows the structure of two base triplets. The yellow nodes are U-A-A base triplets, and the sky blue nodes are C-G-G base triplets.

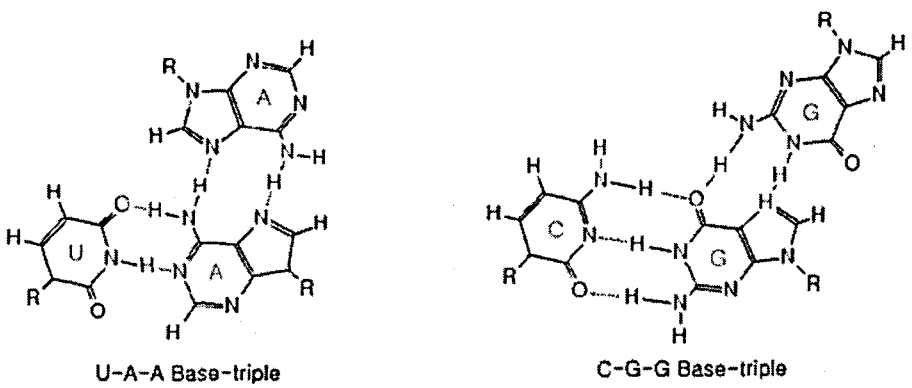

Fig 5. Structure of the U-A-A and C-G-G base triplets at the atomic level; these correspond to the base triplets 1 and base triplets 2, respectively, of Figure 4

The base-triplet structures in Fig 4 agree with those determined experimentally [9, 10].

The merit of our algorithm is its ability to visualize the structure elements of RNA. Programs like Rasmol and Mol-Script can generate the structure of a molecule from the three-dimensional coordinates of its atoms. There are also programs [7] that represent secondary or tertiary structure elements in a plane. However with programs like Rasmol and Mol-Script one cannot easily obtain information about each nucleotide in the RNA and the binding relations between the nucleotides, because these programs represent the structures of molecules at the atomic level. In addition programs that visualize structure elements in a plane have difficulty representing

tertiary structure elements. On the other hand, our algorithm uses the three-dimensional coordinates of the nucleotides to generate secondary and tertiary structures. Hence it produces stereoscopic RNA structures. Moreover it provides not only the configuration of a given RNA molecule but also the bonding relations and types of base pairs between the nucleotides. To the best of our knowledge, this is the first attempt to extract and visualize RNA structure elements from the atomic coordinates in structure databases.

Up to now extracting secondary and tertiary structure elements of RNA from the three-dimensional atomic coordinates has relied upon a substantial amount of manual work. In this study we have developed a set of algorithms for recognizing secondary or tertiary structure elements of RNA in protein-RNA complexes obtained from PDB. Experimental tests showed that our algorithm is easily capable of automatically extracting base-triplet structures and all secondary or tertiary structure elements formed by hydrogen bonding. We expect it to help research on RNA structures, and the regularities in the structure elements discovered should provide useful information for predicting the structure of RNA molecules bound to proteins.

Acknowlegdgements

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

References

1. Tinoco, Jr.: The RNA World (R. F. Gesteland, J. F. Atkins, Eds.), *Cold Spring Harbor Laboratory Press*, 1993, 603-607
2. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22, 1983, 2577-2637
3. Frishman, D, Argos, P.: Knowledge-based protein secondary structure assignment, *Proteins*, 23, 1995, 566-579
4. 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
5. McDonald, I.K. Thornton, J.M.: Satisfying Hydrogen Bonding Potential in Proteins, *J. Mol.Bio*, 238, 1994, 777-793
6. Web-Book Home Page http://www.web-books.com/
7. Han, K., Byun, Y.: PseudoViewer2: visualization of RNA pseudoknots of any type. *Nucleic Acids Res*, 31, 2003, 3432-3440
8. Akmaev, V.R., Kelley, S.T., Stormo, G.D.: Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, 16, 2000, 501-512
9. Du, X., Wang, E.-D.: Tertiary structure base pairs between D- and TΨC-loops of Escherichia coli tRNALeu play important roles in both aminoacylation and editing, *Nucleic Acids Res*, 31, 2003, 2865-2872
10. DNA-RNA structure Tutorials http://www.tulane.edu/~biochem/nolan/lectures/rna/intro.htm