PairAnalyzer: Extracting and Visualizing RNA Structure Elements Formed by Base Pairing

Daeho Lim and Kyungsook Han*

School of Computer Science and Engineering, Inha University, Inchon 402-751, Korea
Kingtiger1@hanmail.net, 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 algorithms were implemented in to a web-based program called PairAnalyzer. 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. PairAnalyzer is accessible at http://wilab.inha.ac.kr/PairAnalyzer/.

1 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 and a program called PairAnalyzer that recognize secondary and tertiary RNA structure elements from the three-dimensional atomic coordinates of protein-RNA complexes obtained from protein data bank (PDB), which provides a rich source of structural data [1]. The structure data were first cleaned up to make all the atoms accurately named and ordered, and no atoms

* To whom correspondence should be addressed. Email: khan@inha.ac.kr

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 279–286, 2004.

have alternate locations. PairAnalyzer identifies hydrogen bonds and base pairs, and classifies the base pairs into one of 28 types [2]. These base pairs include non-canonical pairs such as purine-purine pairs and pyrimidine-pyrimidine pairs as well as canonical pairs such as Watson-Crick pairs and wobble pairs. PairAnalyzer also extracts RNA sequences to integrate them with the data of base pairs. Secondary or tertiary structural elements consisting of base pairs are then visualized for user scrutiny. To the best of our knowledge, this is the first attempt to extracting RNA structural elements from the atomic coordinates in structure databases. PairAnalyzer is intended for analyzing RNA structures. However, it can also be used for analyzing DNA structures since DNA is similar to RNA in hydrogen bonding between complementary bases.

2 Background of Base Pairs and Base Pairing Rules

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 [2] comprising both canonical and non-canonical base pairs. Fig. 1 shows four base pairs.

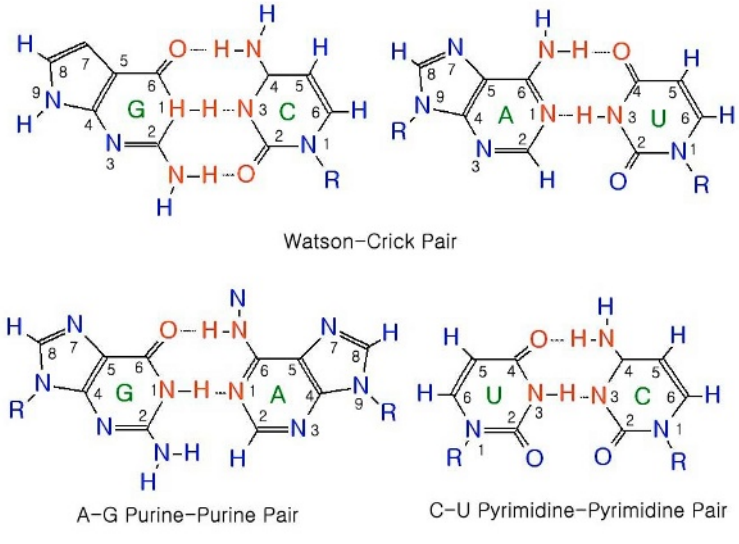

Fig. 1. G-C and A-U Watson-Crick pairs, G-A purine-purine pair, and C-U pyrimidine-pyrimidine pair

A base consists of a fixed number of atoms (see 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 by means of these base pair rules.

3 Algorithms

Our algorithm is divided into two parts. The first part extracts information about secondary and tertiary structure elements of RNA by analyzing data in a PDB file [1]. We use HB-plus [3] 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 second part derives a visual representation 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 and 3 show the framework of the first part and second part, respectively. This section describes the algorithms for the two parts.

3.1 First Part: Extracting Structure Elements of RNA

The first part consists of 5 steps, and the final output is information about the secondary and tertiary structure elements of the RNA.

Step 1: From a PDB file, extract data on all the hydrogen bonds by using HB-plus [3], and record this data in Hydrogen Bonds.

Step 2: Extract the RNA sequence data by analyzing the PDB file, and record it in RNA-SEQ.

Step 3: Extract only those hydrogen bonds that bond one base to another from the hydrogen bond data obtained in Step 1. Record these hydrogen bonds in the Base-Base List.

Step 4: Extract those hydrogen bonds that are involved in base pairing, and classify them into the 28 types by means of the Base pair rules. Record these hydrogen bonds separately in the Base-Pair List.

Step 5: Integrate the sequence data in RNA-SEQ with the base pairs data in the Base-Pair List. Match all the nucleotides in RNA-SEQ to the nucleotides in the Base-Pair List to determine the hydrogen bonding relationships of each nucleotide.

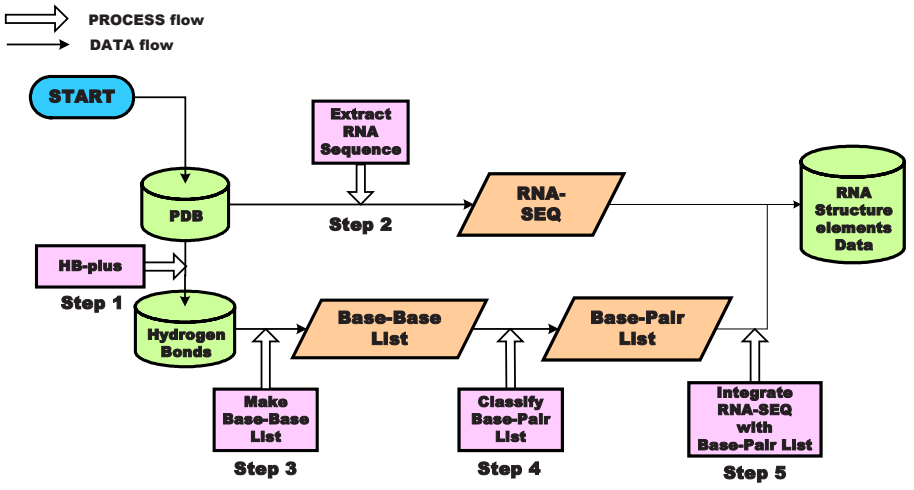

Fig. 2. Framework for extracting base pairs of 28 types and structure elements of RNA from PDB

3.2 Second Part: Visualizing RNA Structure

The 3 steps of the second part are outlined as follows.

Step 1: Obtain the 3D coordinate of every nucleotide by computing the average coordinate values of all the atoms of the nucleotide.
Step 2: Integrate the 3D coordinates of nucleotides with structure elements, and derive the connectivity relation among nucleotides.
Step 3: Represent the structure of the RNA visually by combining the information about structure elements with the coordinate values of the nucleotides.

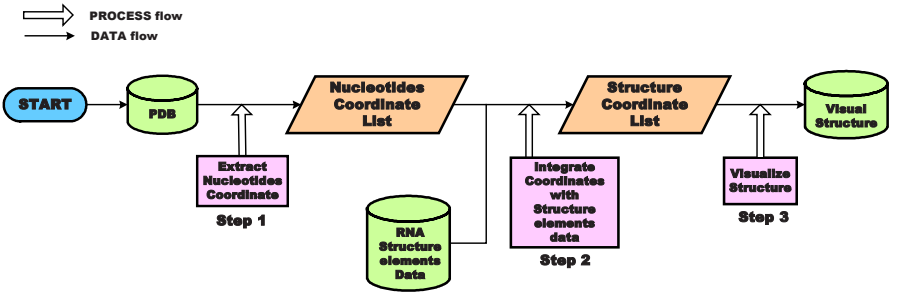

Fig. 3. Framework for visualizing the RNA structure

4 Experimental Results

PairAnalyzer is written in 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. PairAnalyzer takes as input PDB file and HB2 file. As output, PairAnalyzer produces a drawing of RNA structure elements and information about base pairs. Fig. 4 displays the tertiary structure of tRNA (PDB identifier: 1EHZ) derived by PairAnalyzer. Nodes of the drawing 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. The text window in the left shows the information about the nucleotide sequence, base pairs and their types of tRNA.

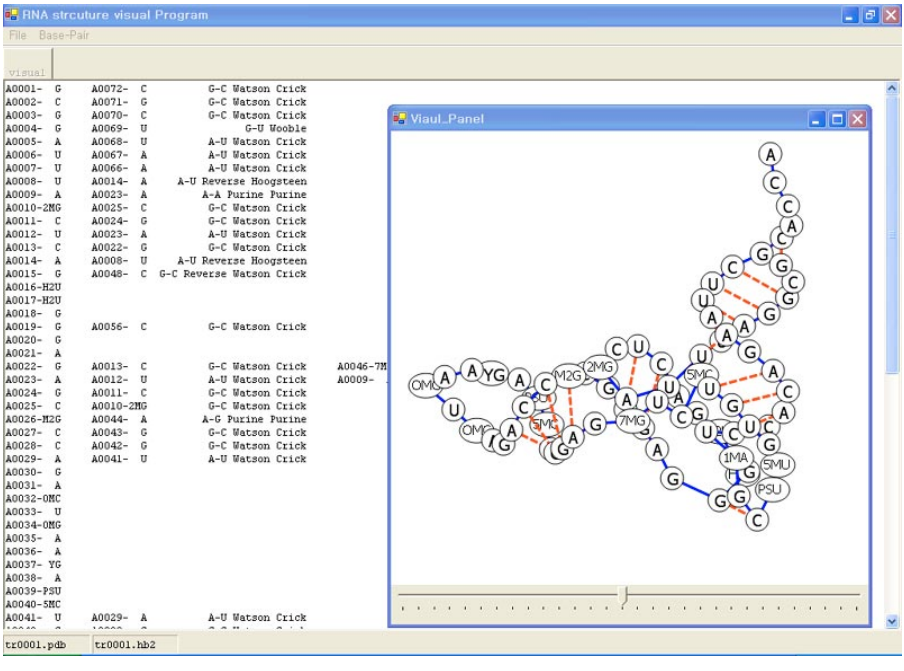

Fig. 4. Interface of PairAnalyzer with the main window and visual panel

Although PairAnalyzer is intended for analyzing RNA structures, it can be used for analyzing DNA structures since DNA is similar to RNA in hydrogen bonding between complementary bases. Fig. 5 shows a small structure of Z-DNA (PDB identifier: 249D) obtained by PairAnalyzer. Z-DNA is a left-handed structure [4]. PairAnalyzer can extract structure elements of left-handed nucleic acids as well as regular nucleic acids because its structure analysis is based on extracting base pairs formed by hydrogen bonds.

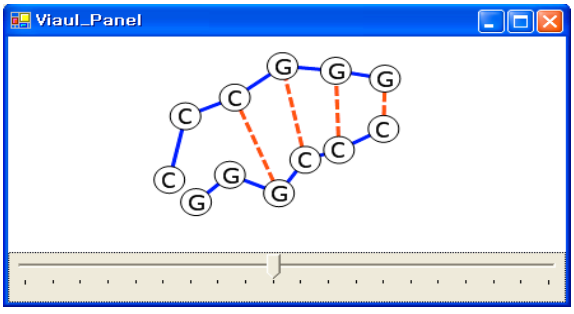

Fig. 5. Z-DNA structure (PDB identifier: 239D) extracted by PairAnalyzer

PairAnalyzer can extract a structure involving multiple RNA stands. The structure shown in Fig. 6 has 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 [5].

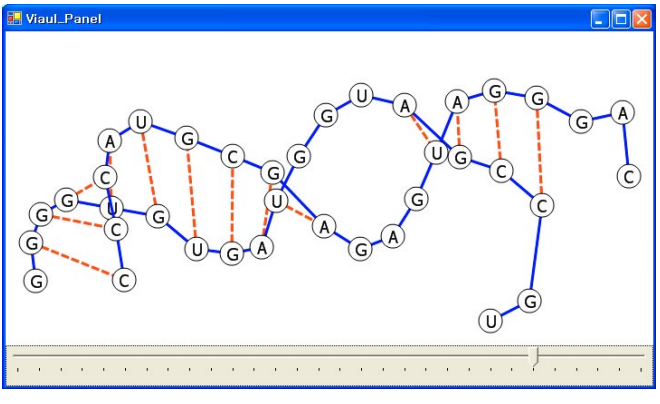

Fig. 6. Structure to have 2 RNA chains (chains M and N), extracted from a protein-RNA complex (PDB identifier: 1DFU)

Programs like Rasmol [6] and Mol-Script [7] can generate the structure of a molecule from the three-dimensional coordinates of its atoms. There are also programs 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.

Fig. 7 shows the tertiary structure of domain V of 23S ribosomal RNA (PDB identifier: 1FFZ) drawn by PairAnalyzer. This structure consists of many nucleotides

and has complex structure elements. If input data in PDB file is given, PairAnalyzer can analyze any structure and extract information about structure elements.

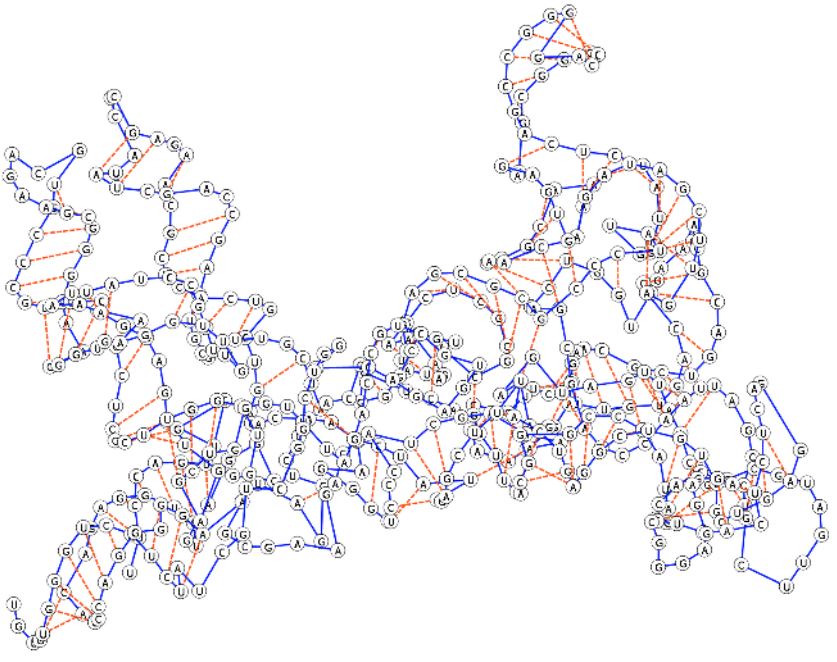

Fig. 7. Tertiary structure of domain V of 23S ribosomal RNA (PDB identifier: 1FFZ) extracted and visualized by PairAnalyzer

5 Conclusion

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. 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. 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.

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

References

1. 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
2. Tinoco, Jr.: The RNA World (R. F. Gesteland, J. F. Atkins, Eds.), Cold Spring Harbor Laboratory Press, (1993) 603-607
3. McDonald, I.K. Thornton, J.M.: Satisfying Hydrogen Bonding Potential in Proteins. J. Mol.Biol. 238 (1994) 777-793
4. Lubert Stryer.: Biochemstry. 4edn. W.H.Freeman, New York (1995)
5. Akmaev, V.R., Kelley, S.T., Stormo, G.D.: Phylogenetically enhanced statistical tools for RNA structure prediction. Bioinformatics 16 (2000) 501-512
6. Roger Sayle : *RASMOL*, http://www.umass.edu/microbio/rasmol/
7. Per J. Kraulis: MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, Journal of Applied Crystallography. 24 (1991), pp 946-950.