Computational Approaches for Investigating Three-Dimensional Structures of RNA

Mohd Firdaus-Raih
School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM-Bangi, Selangor Darul Ehsan.

Abstract
The functions of many complex RNA molecules, including ribozymes and ribosomes, are dependent on their folding into well defined 3-D structures. Intramolecular interactions such as hydrogen bonds have important roles in stabilizing functional tertiary structures of RNA. The 3-D structures of RNA are usually analyzed and annotated through exhaustive visual examination of models generated from X-ray diffraction data. The increasing volume of structural data and diversity of folds have resulted in the need to develop computational approaches to more efficiently and exhaustively analyze as well as annotate 3-D structures of RNA automatically. Approaches which use graph theory have proven to be effective when employed for automatic identification of interaction patterns and structural motifs present in RNA tertiary structures. Hydrogen bonded base interactions such as pairs, triples, quadruples, quintuples and the A-minor motifs are some of the targets which have been used to define the parameters for annotating the 3-D structure of an RNA molecule. In addition to annotation, graph theoretical concepts have also been implemented to represent RNA conformations. This has enabled the comparison of RNA folds and the prediction of novel folds.

Keywords: RNA, 3-D structure, tertiary motifs, graph theory

Diversity of Structures and Functions of RNA
Ribonucleic acids (RNA) participate in numerous key cellular processes. An atomic level understanding of interactions between components of an RNA molecule can contribute greatly towards further understanding of the structural factors and mechanics which enable RNA molecules to correctly fold and carry out their biological function. The RNA molecule is now known to perform a wide spectrum of important biological functions ranging from information storage and transfer (such as in some viral genomes, mRNA and tRNA), to catalytic functions (such as the ribozymes) [1-4], regulatory roles (such as riboswitches) [5, 6] and their long recognized roles in protein synthesis. In order to execute this array of functions, RNA molecules have a wider repertoire of structural variety than the canonical double helical structures commonly associated with nucleic acids, particularly DNA [7].

There is also increasing commercial interest in the use of RNA molecules as molecular scaffolding in nanotechnology [8], or biomedical applications such as drug targets for riboswitches [9] and tRNA [10]. One perhaps shared requirement to achieve these goals is an intimate and in-depth understanding of the three-dimensional (3-D) structure of RNA and the relationships of the 3-D structure to its function. Currently available structures of RNA molecules show that they adopt diverse backbone conformations [11]. This diversity is attributed to the flexibility of the polyribonucleotide backbone as the result of each nucleotide having seven torsional degrees of freedom [12]. Such structural flexibility also makes associating a single native structure to an RNA polyribonucleotide sequence difficult as the possibilities for misfolding are numerous [13]. Several RNA structures are known to be able to self-fold such as small ribozymes while the folding of larger structures such as the ribosome are assisted by protein chaperones [13-15].

The availability of high resolution 3-D structures of RNA from X-ray crystallography has enabled the study of the basic interactions which formed and stabilized the folded RNA molecule. Many of these interactions involve the formation of hydrogen bonds. The base component of a nucleotide can be divided into three edges: (i) the Watson-Crick face which is involved in the canonical Watson-Crick interactions, (ii) the sugar edge on the side of the glycosidic bond and (iii) the Hoogsteen edge (for purines) and CH edge (for pyrimidines) which are on the opposite side of the sugar face. Non-canonical base-base interactions can also occur as a result of hydrogen bonds involving one non-Watson-Crick base edge and was first reported by Karst Hoogsteen [16]. Another type of non-canonical interaction may also occur which however involves the canonical Watson-Crick edges and was proposed by Francis Crick as the wobble-hypothesis [17].
RNA Structural Databases

Databases in molecular biology and bioinformatics have become important resources as data repositories and for data sharing. This importance is reflected in the increasing numbers of databases being developed and made publicly available. There are currently 1078 unique molecular biology databases listed in the 2008 Database Issue published by *Nucleic Acids Research* [18]. From this number, sixteen databases have been classified as databases containing nucleic acid structure data. This number however excludes nucleic acid structural data which are available in larger resources such as the Protein Data Bank which also contain structural data of other biological macromolecules [19].

Structure Repositories. The Protein Data Bank (PDB), now containing over 52 000 structures, remains as the largest one-stop repository for macromolecular structural coordinates available in the public domain (http://www.rcsb.org/pdb/) [19]. Detailed knowledge regarding the atomic structure of an RNA molecule is important in elucidating the mechanisms it employs in performing its biological function. The realization of RNA's crucial roles in biological systems, coupled with improvements in the field of nucleic acid crystallography [20], is also reflected in the PDB [19], which has seen a steady increase of RNA 3-D structures deposited (Figure 1). Currently, over 900 structures solved by X-ray diffraction from a total of over 1400 structures containing RNA chains are available.

The large and diverse collection of macromolecular structures archived in the PDB created a need for dedicated databases and searching methods for nucleic acids and nucleic acid associated structures. Such specific resources will enable existing data to be further analyzed and presented in greater detail to a more specialized audience. Among the sixteen previously mentioned nucleic acid databases which take up this role, the Nucleic Acid Database (NDB) [21] (http://ndbserver.rutgers.edu), a database of 3-D nucleic acid structures, is the most similar to the PDB in terms of available data types. The NDB's contents include database entries which have been sourced from PDB structures in addition to structures which have been directly deposited to the NDB such as those of small nucleic acid structures.

RNA Structural Classification Databases. The SCOR database provides a survey of 3-D motifs in RNA structures sourced from NMR and X-ray data where for some structural classification, the data is further grouped according to RNA function [22]. In addition to RNA structural classification by motifs and function, annotations of tertiary interactions, especially for the larger structures such as ribosomal RNA, were also carried out. The classification curated in SCOR is the result of visual inspection of RNA structures [22].

RNA Base Interactions Databases. Databases with content covering records of interactions occurring in RNA structures, or theoretical interactions which can possibly occur in RNA structures, are also available. Two such resources are: (i) an annotation type approach, the NCIR database [23] and (ii) via a theoretical approach, the NAIL database [24]. The non-canonical interactions in RNA structures (NCIR) database is an extensive annotation of non-canonical base interactions ranging from base pairs to quadruples compiled through a literature survey and is available online (http://prion.bchs.uh.edu/bp_type/) [23]. Entries for the NCIR database include data fields for citation information, the immediate sequence context, a brief common name for the structures in which a particular interaction was observed, a visual representation of the interaction geometry and additional structural information which includes melting point, chemical shifts and specific comments whenever these additional information are available [23]. NCIR is able to provide rapid access to all the RNA structures in which a particular or rare base-base interaction has been

![Figure 1: Annual growth rate of RNA structures in the Protein Data Bank (PDB) from the first RNA structure release (1978) for the database to August 2008. Source: Protein Data Bank - http://www.rcsb.org/pdb.](http://www.rcsb.org/pdb)
observed [23]. However, the capability to explore and document novel interactions which are present in available RNA structures, but which have not been mentioned in literature or missed by visual inspection is still lacking.

Theoretical work in exploring possible base interaction configurations has also been carried out. This was done by first investigating all possibilities of base pair interactions. The outcomes of such a study were the twenty eight base pair possibilities predicted by Saenger [12] together with an additional base pairing possibility presented by Tinoco and co-workers [25] which resulted in a total of twenty nine base pairing possibilities. An excellent example of further work which added to and complemented this foundation, is the library of computed nucleic acid interactions or Nucleic Acid Interaction Library (NAIL) which was made available by the Frankel laboratory at the University of California, San Francisco [24]. The modelling approach used for generating the NAIL database involved systematically computing possible base interactions by matching all possible hydrogen-bond donors and acceptors between bases, followed by evaluating the geometries of each planar configuration [24]. In addition to the four standard bases, Frankel and co-workers also included protonated adenines and cytosines as an interaction component. This in a way filled the gap by accounting for interactions which are novel and may not have been previously reported in literature.

Development of Computational Methods for Analysing 3-D Structures of RNA

Due to the limited numbers of available high resolution RNA structures, much of the early work in exploring RNA structures were actually limited to extrapolating sequence data to structural levels. The earliest molecular model of a complex RNA structure was the transfer RNA model provided by Michael Levitt [26]. Crystallographic structures [27, 28] later revealed the high degree of accuracy and insight contained within the modeled structure where many similarities between the predicted and crystallographic structures were observed such as the arrangement of the four base paired regions into two helical domains. In addition to modeling, the paucity of actual 3-D structures has also led to significant developments of 2-D methods to annotate as well as predict interactions between RNA base components and secondary structures. The work of Gutell et al. enabled the prediction of base pairing interactions in RNA structures to be extracted from sequence information through comparative analyses [29]. These comparative sequence analyses methods have been made accessible via the Comparative RNA Web site (http://www.rna.ccbb.utexas.edu/) and is widely used for secondary structure prediction and representation diagrams.

The first high resolution 3-D structures of the ribosomal subunits [30, 31] have shown that most canonical and even non-canonical interactions have previously been correctly predicted and ‘folded’ into the correct secondary structure interactions. As the number of structures deposited increased, a need arose to develop computational methods for analyzing and comparing RNA 3-D structures. The superficial visual similarity between adenine and guanine; and between cytosine and uracil results in the difficulty of visually scrutinizing RNA structures for specific patterns or motifs. In such a situation, patterns of interactions may not be immediately obvious. This problem is further compounded by the fact that minor changes to relative base orientation may result in different interactions. An additional factor which needs to be considered when implementing a search engine to explore RNA 3-D structures is the relatively larger volume of possible structural equivalences which need to be examined or filtered to achieve an acceptable or operable accuracy margin. This arises from the fact that there are only four types of bases (discounting modified bases) as opposed to twenty amino acids to provide atomic level differentiation of structural components.

Due in part to these factors, the development of pattern searching algorithms to explore arrangements of nucleic acid structural components has become more challenging than for proteins. Secondary structures such as helical regions and interactions such as base stacking and standard hydrogen bonding are of great importance structurally and functionally, however, in the context of a structural search, they may be trivial in the sense that they swamp out more subtle and unusual resemblances. In the late 1990s and early 2000s, several computational methods were developed for RNA 3-D structure analyses [32-35]. These structure analysis programs then led to the need to compare the different structures available with the additional objective of discovering novel substructures and motifs. Graph theoretical approaches have proven to be effective methods for the comparison of 3-D molecular structures and have been used to explore different types of 3-D structures from small molecules to macromolecules such as proteins and nucleic acids.

Graph Theoretical Approaches for Investigating Biological Data

Graph theory is a branch of mathematics which deals with the use of topological representations to study relationships and has been applied to a variety of questions posed by molecular biology ranging from studies on structural data to research on interactions or relationships of data types such as gene expression [36, 37] or protein complexes [38, 39]. These relationships can be further integrated into larger and diverse datasets such as for systems biology type approaches [40]. Graph theoretical approaches which can represent and compare biological macromolecules have been used previously for proteins [41-45], carbohydrates [46] and nucleic acids [47-50].
In graph theory, graphs are mathematical constructs which consist of nodes or vertices related to one another by edges or arcs (Figure 2). These graphs are effectively simple topological representations of what may be complex relationships. Two graphs are said to be isomorphic when there is an exact correspondence between their nodes (Figure 2). Computational procedures which are able to solve this problem are known as subgraph isomorphism algorithms.

Figure 2: The mathematical construct of a graph in graph theory consists of nodes or vertices and arcs or edges which represent the relationships between the nodes. As an example, Graph W can be considered as a subgraph of the topological representation Graph A where W-X-Y-Z (Graph W) is isomorphic to C-D-F-G of Graph A.

Table 1: List of computer programs, the approaches which they implement and the target function of the programs in the general area of RNA 3-D structure analyses.

<table>
<thead>
<tr>
<th>Program</th>
<th>Representative Components</th>
<th>Representation of:</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC-Annotate</td>
<td>Graph representation:</td>
<td>Base interaction geometries</td>
<td>Annotation of RNA 3-D structure</td>
</tr>
<tr>
<td></td>
<td>Node = Nucleotides</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edges = Spatial interactions between pairs of nucleotides</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIMOS</td>
<td>Pseudotorsion angles</td>
<td>RNA conformation</td>
<td>Comparison of RNA conformations.</td>
</tr>
<tr>
<td>RAG</td>
<td>Tree graph representations for RNA structures and dual graphs representations for secondary structures. Node/Edges = multiple nucleotides or base pairs</td>
<td>RNA conformation</td>
<td>Accounting of RNA conformational space and prediction of novel RNA topology.</td>
</tr>
<tr>
<td>NASSAM</td>
<td>Graph representation:</td>
<td>Base interaction arrangement</td>
<td>Annotation of RNA 3-D structure and structural databases, and searching of RNA specific tertiary interactions or motifs.</td>
</tr>
<tr>
<td></td>
<td>Node = Base pseudoatoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edges = Atomic distances between pseudoatoms as vectors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTS</td>
<td>Aligned sections of RNA 3-D structures</td>
<td>Structurally conserved regions in RNA 3-D structures</td>
<td>Comparison of RNA 3-D structures</td>
</tr>
</tbody>
</table>
and structural complexity of RNA crystallographic structures resulted in a requirement for computational analysis packages which could assist in annotating and exploring the atomic interactions within an RNA structure. The idea of applying graph theoretical algorithms for RNA structure comparison, analysis and annotation evolved later than the equivalent methods which had already been used for protein structures and small molecules. The following section looks at computer programs which have been used to analyze RNA 3-D structures or explore RNA conformations and conformational space (Table 1).

Graph Representation of Nucleotide Orientation Geometries: MC-Annotate

One of the earliest works which employed graph theory in the field of RNA structural bioinformatics was done by Francois Major and co workers at the University of Montreal [33]. Their quantitative approach for analyzing RNA 3-D structures was achieved by calculating base-base orientation geometries followed by the compilation of these annotations into a database. In this work, Major and co-workers generated structural graphs via the computer program ‘MC-Annotate’. The structural graphs in this context were representations of nucleic acid structures where the nodes of the graph correspond to the nucleotide components and the graph’s edges corresponded to the nucleotide conformations and base-base interactions. Annotations were added to these graphs with the first set comprising of the atomic coordinates, torsion angles and nitrogen base spatial interactions while the second level consisted of labels which characterized the nucleotide conformations and base-base interactions of which the base type and the relative glycosidic bond orientation for base pairings are two examples. This graph representation of nucleotide-nucleotide interaction geometries will then enable the use of graph theoretical algorithms. MC-Annotate is available as a web program where query structures are submitted via a web PDB file upload feature (http://www-lbit.iro.umontreal.ca/mcannotate-simple/). MC-Annotate is able to annotate nucleotide interactions in RNA structures. However, its default features are not capable of searching for a specified 3-D pattern or for searching and annotating multiple structures such as a database. Therefore, this program appears to be intended for examination and annotation of individual structures as opposed to the capability to annotate and contrast the content of large databases.

Representation of RNA Conformation Using Pseudo-torsion Angles

PRIMOS (Probing RNA structures to Identify Motifs and Overall structural changes), a computer program implemented in Perl by the Pyle laboratory at Yale University, compares representations of pseudo-torsion angles of RNA structures termed as ‘RNA worms’ (Figure 3) [47]. The RNA worm representation of RNA 3-D structure enabled the comparison of different backbone conformational states of closely related molecules. PRIMOS does not directly locate occurrences or track changes of base interactions in RNA structures. Investigation of such base interactions are done by manual visualization after first identifying conformational changes of interest between RNA worm representations of different structures.

While PRIMOS has been demonstrated to be able to detect structural motifs such as the bulged-G / loop-E motif, it is not expected to be able to detect base interaction changes which do not affect the RNA worm representation significantly. An example of such possible interactions are in highly conserved folds with diverse base content, where a base pairing change results in the isostructural maintenance of the phosphate backbone and therefore does not result in a detectable difference in the RNA worm representation. Despite this fact, such a tool remains a powerful resource for comparing the overall folding of RNA 3-D structures.

A companion program of PRIMOS, aptly named as COMPADRES (Comparative Algorithm to Discover Recurring Elements of Structure), was developed to enable the discovery of recurring conformational motifs [51]. COMPADRES searches for potential motifs by employing the RNA worm representation from PRIMOS and comparing the RNA worms for recurrences covering at least five nucleotides. As a result, COMPADRES is able to compare the conformational diversity present within a specific dataset or database. The major limitation of COMPADRES lies in the fact that motif discovery is limited to those containing sequential nucleotides.
Graph Representation of RNA Secondary Structure Topology as ‘RNA Worms’

The PRIMOS RNA worm [47] can also be seen as a map of the topology for the RNA molecule it is representing. This approach basically generates a representation of the topology for the input RNA structure and not much else. Although the use of a companion program such as COMPADRES has enabled such an approach to be utilized for discovering recurrent structural motifs, they are unable to explore the myriad of possibilities that a strand of RNA can fold into to become a complex RNA 3-D structure without having first been supplied with a known motif.

RNA secondary and tertiary structure consists of hierarchical and modular substructure components which fold independently [52]. Therefore, methods which can compute and explore the diversity of modular RNA topologies can greatly complement conformation-based comparison approaches such as PRIMOS [47] and COMPADRES [51]. In 2003, Tamar Schlick and co-workers reported the use of graph theory to enumerate tree graphs which can represent the repertoire of possible RNA secondary structures [48]. The outcome of this work was integrated into the ‘RNA As Graphs’ or ‘RAG’ database (http://monod.biomath.nyu.edu/rna/rna.php).

Schlick and co-workers [48] enumerated tree graphs which represented components of RNA secondary structure topologies where the nodes and edges can both represent multiple nucleotide bases or base pairs. The enumeration of tree graphs were first reported by the Victorian era mathematician Arthur Cayley and was used for studying the extensions of hydrocarbons such as alkanes [53]. The representative graphs generated by Schlick and co-workers [48] were further subdivided into two where tree graphs represented RNA trees (or structure) and dual-graphs represented any secondary structures. Dual graphs are non-planar graphs which may include loops whereas tree graphs are considered planar relationships. In order to facilitate their representations, several rules were used to identify the components which can make up their representation scheme [48].

Using a combination of tree and dual graphs, Schlick and co-workers [48] were then able to enumerate their representations in order to derive and describe possibilities of RNA topologies which also include novel topologies. Their graphical enumeration has shown that RNA topology space is very much smaller than sequence space and therefore makes the effort of searching for novel RNA folds plausible. The proof of their approach was presented in the form of their method’s capability in representing an existing type of RNA topology from which they can then differentiate novel topologies currently not available in existing structures.

Representation and Search of Base Orientation Patterns: The NASSAM Program

NASSAM (Nucleic Acid Search for Substructures and Motifs) is a graph theoretic application which implements the Ullmann subgraph isomorphism algorithm [54] for comparing pseudo-atom representations of RNA base orientations [49]. The concept behind the pseudo-atom representation used in NASSAM is a nucleic acid specific evolution of the previously mentioned ASSAM program. Each of the four RNA bases are represented by two pseudoatom vectors consisting of four pseudoatoms; where one pseudoatom is the start node and another pseudoatom serves as the end node as illustrated in Figure 4A.

The NASSAM base representations are labeled graphs in which the pseudoatoms are the nodes of the graph, and the distances between them are the graph’s edges. As an example, inter-vector distances can be used to define a query pattern resulting in matrices (Figure 4B,4C) which provide information on the orientation of the bases to each other (Figure 4A). The matrices for the search and target structures are used by the Ullmann algorithm to search out subgraph isomorphisms which represent the matching of the orientation of bases between the query and the search structure. Additional pseudo-distances were incorporated into the matrices to ensure that the midpoints of the pseudo-atom vectors for a base were close enough to each other that they were on the same base. These constraints were included in all the patterns and were designed to avoid possibly erroneous searches which utilized the start node on one base and the end node of a different base. Throughout the work reported by Harrison et al [49], midpoint to midpoint distances internal to a base were constrained to 1Å (Figure 4B,4C).

In addition to the input data, input parameter requirements for running a NASSAM search include: (i) a distance tolerance parameter, (ii) distances from which vector types to utilize for a search, (iii) sequence order parameter and (iv) a restraint for the distance from the hit to the nearest heteroatom [49]. NASSAM searches can be made independent of sequence order with no distance restraints set for the nearest heteroatom. The vector types parameter can be set in the input pattern file which will be read by the NASSAM program as the default with the possibility of user intervention prior to a search. As an example, a search for a triple pattern may have been set by default to use SS, SE, EE and MM vectors (S = Start, E = End, M = midpoint); which a user can have the option of dismissing or adding a vector type to utilize. This results in a search which will employ less vector types such as a search with SS, SE and MM distances only.

The distance tolerance parameter is a value which sets the amount of deviation from the distances supplied in
Figure 4: Pseudoatom representation of relative base orientations used by NASSAM. (A) Example of a base triple composed of a guanine (Base 1), a cytosine (Base 2) and another guanine (Base 3). The pseudoatom nodes used to set the distances for the pattern matrix have been marked with $S_x$, $E_x$ and $E_y$ while the distances between these nodes have been marked with arrows. (B) An example set of vectors and their corresponding distances (in Angstroms) which describe the GGC triple orientation in panel (A). (C) The corresponding pattern matrix file built from the vectors (distances X10) defined in panel (B) for the triple pattern shown in panel (A). As an example, the $S_xS_x$ distance between Base 1 and Base 2 is indicated as 9.2 Å in panel (B), and is marked 92 under the SS column where Base 1, Node X and Base 2, Node X intersect and the SS column where Base 2, Node X and Base 1, Node X intersect in the matrix in panel (C).

The distance tolerance parameter has been applied for RNA structure searching in two ways. The first involves the use of a low percentage value which gives the best balance between precision and recall values, which can be used for annotation of known RNA structural motifs or structural elements in a new structure. The second approach uses high distance tolerance values exceeding 50%, which results in a ‘fuzzy’ search where the query pattern is able to represent a larger number of possibilities including orientations which vary from, or are derived from the original input pattern. This feature of NASSAM has been used for the discovery of novel base orientations. The authors of ARTS have written in the ‘Introduction’ section of their report that both NASSAM and PRIMOS were unsuitable for detecting new motifs which had not been pre-specified [55]. This statement is inaccurate for NASSAM because the use of fuzzy searching has enabled the discovery of base orientations which have not been pre-specified in the query pattern [56]. This in part proves NASSAM’s suitability for use not only in RNA structural annotation but also as a tool for the discovery of novel structural interactions.

Alignment of RNA Tertiary Structures: ARTS

ARTS or ‘Alignment of RNA Tertiary Structures’ was reported as a novel method that was able to align nucleic
acid tertiary structures and thus enabled the detection of common substructures [55]. The ARTS approach begins by superposing the largest number of phosphate atoms for one structure on to the phosphate atoms of the second input structure. This initial seed match is then extended to coinciding base pairs and unpaired nucleotides, then scored and ranked with the highest scoring matches being reported. The second stage after seed match extension was achieved by finding the maximal matches in a bipartite graph. The output transformations were further refined by least squares fitting [57].

The developers of ARTS attempted to fill a void in *a priori* detection of novel substructures in nucleic acids via pair-wise structure comparison. This pair-wise approach was further extended to conduct whole database alignments and was reported to be highly efficient and suitable for searching through the contents of large 3-D structure databases such as the PDB. An exhaustive all-against-all comparison of 770 RNA 3-D structures was conducted using the ARTS method as a proof of concept [55]. However, no novel tertiary motifs were reported.

**Annotation and Discovery of Structural Motifs in RNA Structures**

The increasing availability of RNA 3-D structures has led to a need for computational methods which can efficiently and accurately analyze and compare these structures. The use of computational approaches, such as graph theoretical methods, has proven to be successful in addressing these needs. Despite the reported successes, several gaps still need to be filled particularly with regard to methods for discovering novel motifs. The capability to describe structural interactions and search for them computationally leads to the outcome of collections of uniformly annotated RNA structures. This aspect of standardization in annotation can be an additional tool for discovering novel motifs by comparing and contrasting annotated RNA structures.

**Acknowledgements**

Many useful comments from Prof. Peter Artymiuk (Molecular Biology and Biotechnology, University of Sheffield) and Prof. Peter Willett (Information Studies, University of Sheffield) have contributed towards the preparation of this manuscript.


