An RNA Secondary Structure Analyser for RNA Molecule Design

CPSC 448 Proposal

Version 1.4, 2002-11-12

Farheen Rawji, frawji@rational.com

Supervisor: Holger Hoos

Abstract. Ribonucleic acids (RNA's) are one of the most important biological polymers active in living organisms. The functions of many classes of RNA's are determined primarily by their structures. The design of RNA molecules of known structures hence has applications in biological and biochemical research, as well as in industrial applications of this research. There are also applications in DNA computing and molecular nanotechnology. In this project, I will outline the parameters and functionality for an *RNA Structural Analyser* which will be used to generate statistical information on known RNA's. This information can then be used to generate RNA's with biological features, using contemporary RNA design algorithms.

1. Introduction

1.1 RNA Molecule Structure

An RNA (Ribonucleic Acid) molecule is composed of a string of 4 different nucleotides (base molecules that make up RNA or DNA), namely A, U, G and C (in the case of DNA the bases are A, T, G and C). These bases are strung together to form what is termed the primary sequence of the RNA molecule. The primary structure thus can be conceptualised as a string of beads, where the beads are just the individual nucleotide bases as shown in Figure 1. One end of the molecule (referred to as the 5' end) can be distinguished from the other end (the 3' end) and by convention, the primary sequence is usually written 5'-3' from left to right.

This string of nucleotides can chemically fold on itself via the specific pairing of A with U and G with C (sometimes other pairings also occur, e.g. G-U pairs) to form the secondary structure, which consists of a distinct set of structural elements such as loops and stems. Some of the conventional secondary structure elements are:

- Stems.
- Hairpin Loops, Tetraloops
- Interior Loops
- Multibranched Loops.
- Bulges.
- Tetraloops.
- Pseudoknots
- U-Turns.

Tetraloops are an especially stable form of 4-residue hairpin loops. Pseudoknots are formed by a pairing of a single-stranded section of the RNA to the unpaired residues in a hairpin loop to form coaxially stacked stems. Hence a *pseudoknot-free* structure is one which does not have any two base pairs (i , j) and (i' , j') where $i < j < i' < j'$(see Figure 3 for details). Some of these secondary structure elements are depicted in Figure 2. The tertiary structure is created by a further set of interactions of the secondary structure elements.

A good summary of RNA secondary and tertiary structures is given by Westhof and Auffinger [10].

NOTE: Figures 2 and 3 have been taken from course notes for BIOC 410 at UBC, Winter 2002, Prof. G.A. Mackie

Figure 1 Primary structure of an arbitrary RNA molecule. The particular sequence of bases distinguishes one molecule from another.

5'.....AGUUUGGCGACU.....3'

Figure 2 Secondary structure of an arbitrary RNA molecule. There may be any combination of these structures in a molecule, the particular structures being determined by the base pairing interactions within the primary sequence of the molecule. The black bars indicate A-U, G-C or G-U pairs. The open bars represent unpaired nucleotides.

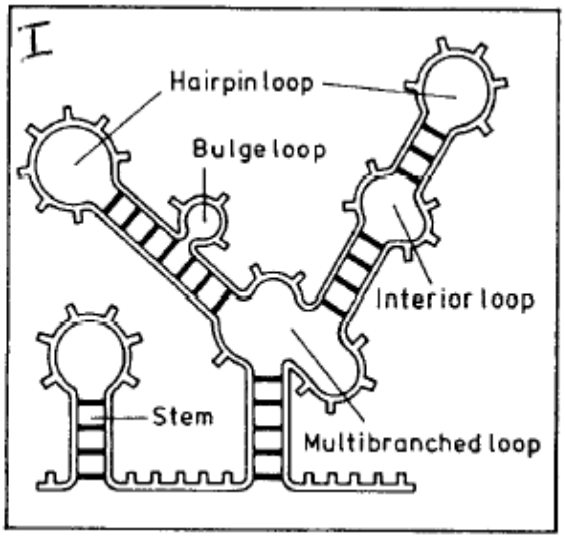
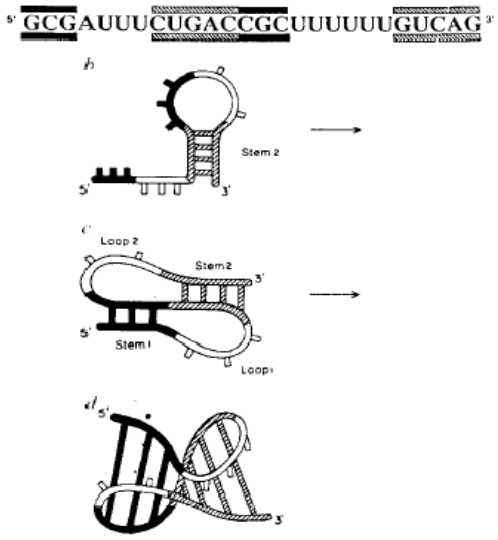


Figure 3 Schematic representation of the formation of a pseudoknot structure.



1.2 Determination of RNA Molecular Structure

RNA molecules act as intermediaries in the process of converting the ubiquitous DNA code into cellular proteins. RNA molecules are also important for a number of other roles they play in cellular mechanisms, such as the enzymatic activity of catalytic RNA molecules called ribozymes. Functionality of RNA molecules is crucially dependent upon their structure.

Traditional methods of determining structure rely on biological and biochemical methods. The primary structure or sequence of nucleotides in an RNA molecule can easily be found by biochemical methods such as the dideoxy method [1]. However, determination of secondary and tertiary structure is a complex, time-consuming process, producing - quantifiably and qualitatively - limited results. Hence, computational methods for predicting the secondary and tertiary structures of these molecules has wide applicability for research work in the biological and biochemical fields as well as for industrial applications in these fields.

One such computational approach involves determining the most likely folding of a known sequence into a secondary structure. This is referred to as the **RNA Secondary Structure Prediction Problem**: given a primary RNA sequence, determine the most viable secondary structure that this sequence will fold into. Various algorithms have been implemented to produce such optimal foldings, based on calculations of minimum free energies of formation of the secondary structures. Implementations that solve this problem for the pseudoknot-free case in $\theta(n^3)$ time depend on dynamic programming algorithms such as that of Zuker [2]. Most implementations have been thus far limited to determination of *pseudoknot-free secondary structures*. Examples are the MFOLD [3] and Vienna [4] folding algorithms.

1.3 The Problem of RNA Secondary Structure Design

Another approach to structure determination relies on consideration of the application for which the RNA molecule is needed. Since functionality depends on structure, the goal is to create molecules of specific secondary and tertiary structures. The ability to create RNA molecules based on structural considerations (molecules with predetermined biological functionality) would be of tremendous value in drug design as well as for other biological and industrial applications. Another application is in the field of DNA self-assembly computation [5] as well as in molecular nanotechnology [6].This approach is formalised as the **RNA Secondary Structure Design Problem**: given an RNA secondary structure, find an RNA strand that will fold into this structure. The algorithm for solving this problem would have as input a given secondary structure of an RNA molecule, and would output a predicted sequence that will fold into that structure. Once the sequence is known, it is possible to create the RNA molecules using established laboratory techniques.

[Some of the significant aspects of this approach that relate to my project have been detailed in this section.]

One solution to the RNA secondary structure design problem is provided by Hofacker et al. [7], the implementation for which is included in the Vienna RNA Secondary Structure Package [4]. A more recent stochastic local search (SLS) implementation which has been published by Andronescu et al. [8], shows substantially improved performance.

The performance of the SLS algorithm for secondary structure design (*RNA-SSD*) implemented by Andronescu et al. [8] was tested against those of other algorithmic implementations, such as the Vienna algorithm [4] (please refer to the Andronescu et al. [8] paper for further details). These tests were conducted using test structures for which the primary and secondary structures were known. An important consideration for these tests is whether the structures being tested are biologically plausible, since the eventual application of the algorithm, for the most part, is towards creating RNA's that have biological structures and functions.

There are two ways of obtaining such test structures: one is to retrieve them from biological molecule databases, such as the Ribosomal Database Project [9] and the other is to generate them computationally. Relying solely on databases for test structures is problematic since they do not provide large enough numbers of test molecules. Another consideration is the ability to test structures that have certain well-defined features, for which the algorithm is to be specifically tested. Hence, it may be desired to test for a set of 'imaginary' molecules that have three stems, two hairpin loops and one multibranch junction. This is a specifically designed test constraint which cannot be derived from biological databases (except for maybe a lucky coincidence!).

The alternative to obtaining molecules from a database is to simulate the structures using a structure generating algorithm. Such an *RNA Structure Generator* has been implemented in the research carried out by Andronescu et al. [8] at the BETA lab at UBC. The RNA Structure Generator is a simple random generator for pseudoknot-free RNA secondary structures. One criterion of a 'good' set of test structures is that it should represent the biological features present in naturally occurring RNA molecules. This is important since one purpose of designing RNA molecules is to design molecules that have biological functionality as well as to create molecules that have industrial applications. RNA molecules have many features in common with DNA molecules, the biological features of which are exploited in DNA tiling experiments and molecular nanotechnology applications. Thus, it is important to know what the salient features of biological RNA's are, in order to use this to build biologically representative RNA's using the RNA Structure Generator.

2. Designing an RNA Secondary Structure Analyser to Discover Salient Features of RNA Molecules

The objective of my directed studies research for CPSC 448 will be to design an **RNA Analyser** to discover and classify the salient secondary structure features of biological RNA molecules. This information will then be used to create biologically representative RNA's using the RNA Structure Generator, which can then be used as test subjects in RNA Structure Design problems.

2.1 Functionality of the RNA Analyser

The RNA Analyser will take as input a format representative of the secondary structure of an RNA molecule (see section 2.1.1 for details) and use this to generate a summary of the **secondary structure features** of the molecule. The list below gives a set of features the Analyser will determine:

NOTE: This set may be slightly modified as the project progresses.

a. **Number of Stems** (e.g. 3 stems).

b. Number of Loops (e.g. 3 hairpin loops, 1 bulge loop, 1 interior loop, 1 multibranched loop).

c. **Length** (e.g. Stem 1 is 4 base pairs long).

d. **Positions of structural elements within RNA molecule:** the bases in the RNA Molecule of length n are numbered from position 1 at the 5' end to position n at the 3' end. (e.g. Stem 1 is formed by the bases between positions 2 - 5 and 12 - 15).

e. **Number of arms and unpaired bases for all loops** (e.g. Multibranched loop with 3 arms and 7 unpaired bases, bulge with 2 unpaired bases).

f. **Occurrences and structure of all pseudoknots** (e.g. pseudoknot between bases 63-65, 87-89, 77-80 and 102-105 on hairpin at Stem 3).

2.1.1 Accepted Input Formats

The RNA Analyser will accept commonly used formats for representing RNA secondary structure elements. Two commonly used formats are:

i. **Dots and Parentheses:** where dots represent unpaired bases and parentheses represent paired bases. For example, a simple stem ending in a hairpin loop such as Stem 1 with its adjacent unpaired single-stranded region in Figure 2 would be represented as:

.((((.....)))).....

ii. **Specific Base interactions:** for example, if the sequence for the above stem was AGCGCGUUAUUGCGCUGGAA then the representation would begin as:

1 A 0

2 G 15

...

15 C 2

...

20 A 0

2.1.2 Determination of Statistical Models of Structures

Using the feature summary generated for each molecule, the Analyser will compile these results statistically in order to create distributions of features over a number of classes of molecules. For example, the RNA's may be classified into the following general 'types' and for each type, the features with means and standard deviations can be found:

Some examples of General Classes of RNA's

I. Messenger RNA (mRNA).
II. Ribosomal RNA (rRNA):
    a. 23S rRNA
    b. 5S rRNA , etc...
III. Transfer RNA (tRNA).
IV. Ribozymes:
    a. Group I Introns
    b. Group II Introns, etc...

V. Small RNA Molecules (snRNA):
    a. U1 snRNA's
    b. U2 snRNA's, etc...

A feature summary for U-RNA's may be, for example:

- Mean number of stems/ molecule = (3 ± 0.07);
- Mean length = (205 ± 3) nucleotides.
- GC content of stems, loops, etc.

Hence, the Analyser will generate information about the general classes of RNA molecules that can then be used to learn statistical models of RNA structures. This information can then be applied to generate sets of biologically realistic structures using the RNA Generator.

2.1.3 Simple Modifications of Structures

The Analyser will have functions for automatically removing pseudoknots, as well as other simple modifications of existing structures.

2.2 'Raw Data' for the Analyser

The raw data for the Analyser consists of the set of RNA's that are input to the Analyser for analysis. These RNA's can be found from biological databases, such as the RDB, as described earlier [9]. Finding other such databases of RNA molecules is a part of this project.

3. Research Objectives

The objectives of this research project are:

1. Design and implementation of a programme for analysing RNA secondary structure.
2. Application of this programme in the context of characterizing secondary structures of biological RNA's.
3. Application of this programme in the context of identifying features that determine the hardness of secondary structure design.

4. **Research Plan**

The project will have the following tasks with corresponding milestones and deliverables:

NOTE: Intermediate deliverable documents such as 'Test Cases' may be delivered as informal notes or email correspondence to the instructor.

TASK	MILESTONE	ASSOCIATED DELIVERABLE
1. Literature search of current methods of RNA secondary structure classification, including algorithms for secondary structure generation. **NOTE:** it is assumed that the literature search will be a continuous process during the entire course of the project in order to incorporate new information as and when needed.	Report findings by **Nov. 22nd, 2002**	**~5 page literature search.**
2. Create Project Web Site with links for research findings and literature review. (Update this site as needed during course of project).	**Due: Nov. 15th, 2002**	**Project Web Site (first draft)**
3. Design of analyser (including statistical analysis to be incorporated in program).	Complete design by: **Nov. 30th, 2002**	**Document of functional and non-functional requirements**
4. Iterative implementation and testing of the analyser.	Design Prototype with minimal functionality (e.g. analyser accepts number of stems and loops and generates statistics based on this information) **Due: Dec. 16th 2002**	**Prototype source Code and demo**
	Design Test cases **Due: Dec. 20th 2002**	**Test Cases doc.**

	Complete implementation by: **December 31st, 2002**	**1. Complete Source Code and associated documents (e.g. test cases, results).**
	Summarize results by: **Jan. 6th, 2003**	**2. Research paper summarizing results for this section.**
	Present results by: **Jan. 10th, 2003**	**3. Demo of RNA Analyser with presentation of research findings**
5. Add RNA Analyser and relevant project information to the website.	Deployment by: **Jan. 20th, 2003**	**Project Website with research paper, associated literature links and source code**
*** See Endnote** 6. Develop sets of characteristics of biological RNA's using the analyser.	Deliver characteristics by: **Feb. 20th, 2003.**	**Review Document**
7. Apply results to identify features that determine hardness of secondary structure design.	Deliver feature list by: **March 15th, 2003.**	**Feature List**
8. Use results of the study to improve the RNA Generator.	Deliver results by: **April 1st, 2003.**	**Results Summary**
FINAL: Write up project in a review paper for submission with the code and website. **Present research to BETA Lab.**	Final Project Deliverables by: **April 30th, 2003.**	**1. Final Research Report** **2. All source Code** **3. Project Website** **4. Research Presentation**

**** Endnote:***

The stages of the project have been well-defined for Term 1 and for Term 2 they remain to be elaborated in a meeting with the instructor at the beginning of Term 2.

References

1. Sanger,F., Nickler, S., Coulson, A.R., *DNA Sequencing with Chain-Terminating Inhibitors*, Proc. Natl. Acad. Sci., USA, 74: 5463 – 5467, 1977.

2. Lyngso R.B., Zuker, M., Pederson, C.N.S., *An Improved Algorithm for RNA Secondary Structure Prediction,* BRICS Report Series, 1999, RS-99-15, Department of Computer Science, University of Aarhus.

3. MFOLD: http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi

4. Vienna Folder: http://www.tbi.univie.ac.at/cgi-bin/RNAfold.cgi

5. Winfree, E., *Algorithmic Self-Assembly of DNA: Theoretical Motivations and 2D Assembly Experiments,* Journal of Biomolecular Structure and Dynamics, Conversation 11, Issue #2 (2000), 263-270.

6. Reif, J. H., *DNA Lattices: A Method for Molecular-Scale Patterning and Computation*, Computing in Science and Engineering, 2002, January/February: 32- 41.

7. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, M., Tacker, M., Schuster, P., *Fast Folding and Comparison of RNA Secondary Structures*, Chemical Monthly, 1994, 125: 167-188.

8. Andronescu, M., Fejes, A.P., Hamze,F., Hutter, F., Hoos, H., Condon, A., *A New SLS Algorithm for RNA Secondary Structure Design,* Department of Computer Science, University of British Columbia, 2002.

9. Ribosomal Database Project: http://rdp.cme.msu.edu/

10. Westhof, E., Auffinger, P., *RNA Tertiary Structure,* Encyclopedia of Analytical Chemistry, R.A. Meyers, 2000: 5222 – 5232.