A Heuristic Approach to RNA Secondary Structure Prediction

Evelyn Stiller
Gregory Riccardi
Department of Computer Science
The Florida State University
Tallahassee, Florida 32306-4019 USA
stiller@cs.fsu.edu riccardi@cs.fsu.edu

Abstract- The overall research context for this endeavor is the creation of an object-oriented representation scheme to facilitate phylogenetic analysis. As an initial and non-trivial subtask of this undertaking, we recognize the need to accomplish sequence alignment preceded by secondary structure prediction due to the length of the RNA strands with which we are working, namely 18S ribosomal RNA (rRNA). Secondary structure prediction is the focus of this paper. Though this topic has been addressed for the past two decades, our efforts concentrate on the introduction of heuristic techniques to make tractable an otherwise intractable problem. We introduce an object-oriented representation scheme to facilitate the utilization of biologist expertise and phylogenetic context into the arrival of secondary structures.

1 Background and Motivation

It has been the case for several years that the production of genetic sequence information has far out paced its analysis [4]. Automated techniques have been developed to facilitate some forms of analysis, but tend to be exhaustive in nature, restricting the utility of such software to relatively small genetic sequences [2]. Because the biology laboratory with which we are collaborating uses 18S ribosomal RNA (rRNA), which typically contains in excess of 1,800 nucleotides [3] (the basic building blocks of DNA and RNA), we seek to refine these analytical techniques to accommodate these genetic sequence lengths.

Needleman and Wunsch [6] are credited with establishing the definitive, recursive sequence alignment approach. Others have adapted this algorithm to include thermodynamic restrictions and thereby streamline the processing somewhat [7]. Davidson [1] surveys sequence alignment/homology search algorithm variations, noting efforts to reduce CPU time and memory requirements. These do not offer a significant change in the order of magnitude of computation, which we hope to accomplish with the possible exception of Martinez [5].

In phylogenetic analysis one seeks to align two or more genetic sequences to deduce the organism's place in evolutionary history, according to some paradigm such as cladistics. The evolution of rRNA frequently entails the insertion or deletion of relatively long series of nucleotide sequences which disrupts traditional alignment algorithms. However, when one views rRNA in terms of its secondary structure these evolutionary changes are more easily recognized.

The secondary structure of a genetic sequence is the thermodynamically stable pairing of nucleotides based on the results of Watson and Crick, who found that in RNA the nucleotides Adenine and Uracil pair and that Guanine and Cytosine pair. These base pairs contribute to a thermodynamically stable molecule, whereas unpaired nucleotides contribute to instability. The process of arriving at a globally optimal secondary structure can be considered algorithmically similar to establishing an alignment with the exceptions that the search takes place within the strand rather than between strands and that rather than seeking matching nucleotides the Watson-Crick partner is sought.

2 Object-Oriented Design and Heuristic Approach

Single stranded RNA such as rRNA folds back unto itself, forming helical areas interspersed with unpaired, single-stranded areas. The helices are formed when Watson-Crick complementary nucleotides are paired in addition to Guanine and Uracil pairs and occasional odd pairs (other than the previously mentioned). Out of

these single and double-stranded regions six generally recognized secondary substructures exist. These substructures may be alternately named elsewhere, but follow the terminology of Gouy [2] as 1) Helix, 2) Bulge, 3) Hairpin curve, 4) Internal loop and 5) Multiloop, and 6) External single-stranded regions. These are given informal as well as formal treatment elsewhere [8, 2]. It is hoped that this representation scheme will not only facilitate the folding algorithm, but also support the integration of phylogenetic context.

Initial analysis of the secondary substructures suggests that they serve ideally in an object-oriented design, where the secondary structure can be represented, very effectively, as a *Set* of interconnected Multiloops, prefixed or suffixed with External single-stranded regions. The Multiloops are comprised of a List of Helix and *Single-stranded region* pairs. The Helices are then comprised of Double-stranded regions, beginning and ending in Multiloops or Hairpin curves, thus providing interconnectivity between Multiloops. All of the substructures previously mentioned, with the exception of two, are thus incorporated in the design, with the remaining two being encompassed by the Double-stranded region. Figure 1 illustrates the relationship between the object classes and the secondary structure of RNA.

The above design entails the introduction of several objects that serve as components to the originally listed substructures, and two aggregating primitives, namely list and sets, in order to make the design complete. The components are 1) Single-stranded regions, 2) Double-stranded regions and 3) Branch-points. These added components decompose into very straight forward definitions, maintaining the simplicity of the design. A single-stranded region is simply a series of nucleotides. Since it is customary in biology to sequentially number each nucleotide in an RNA strand, a Single-strand can simply be represented as a beginning and an end point. In a similar manner, the Double-stranded region consists of two of these end-points, with the refinement that these regions may have Bulges and Interior-loops interspersed, thus incorporating the two previously unmentioned substructures. The two strands of the Double-stranded region are assumed to be Watson-Crick pairable, with the exclusion of the nucleotides belonging to Bulges or Interior loops. The final substructure component is the Branch-point, which is the starting point for helices, requiring the sequence number of two nucleotides.

Presently,in addition to refining the class heirarchy, our efforts are directed toward finding efficient methods for building secondary structures based on previously derived foldings, and to find the translation method that best utilizes such an encoding for facilitating new RNA secondary structure prediction. We must then establish means to assess the quality of the derived folding.

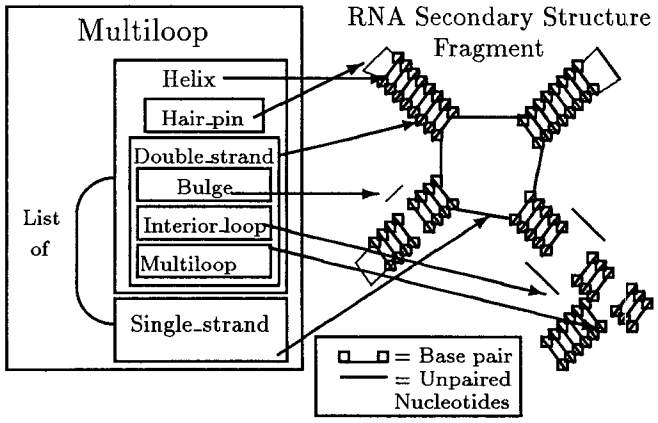

Figure 1: Object heirarchy related to an RNA secondary structure fragment.

References

[1] Dan Davidson. Sequence similarity (homology) searching for molecular biologists. *Bulletin of Mathematical Biology*, 47:437–474, 1985.

[2] Manolo Gouy. Secondary structure prediction of RNA. In M. J. Bishop and C. J. Rawlings, editors, *Nucleic Acid and Protein Sequence Analysis: a Practical Approoach*, chapter 11, pages 259–284. IRL Press, Oxford, 1987.

[3] Won R. Kim and Lawrence G. Abele. Molecular phylogeny of selected decapod crustaceans based on 18S rRNA nucleotide sequences. *Journal of Crustacean Biology*, 10:1–13, 1990.

[4] Arthur M. Lesk. *Computational Molecular Biology, Sources and Methods for Sequence Analysis.* Oxford University Press, Oxford, 1988.

[5] Hugo M. Martinez. An RNA folding rule. *Nucleic Acids Research*, 12:323–334, 1984.

[6] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[7] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal of Applied Mathematics*, 45:810–825, 1985.

[8] David Sankoff and Joseph B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, London, 1983.