

Qualitative Modeling of RNA Structure

Kyungsook Han and Andrew Gelsey

kshan@cs.rutgers.edu gelsey@cs.rutgers.edu

Department of Computer Science

Rutgers University

New Brunswick, NJ 08903

U. S. A.

Abstract

Determining the folding structure of an RNA molecule from its underlying linear sequence is a complex problem involving both spatial reasoning and the use of knowledge of chemistry and biology. Most research in qualitative physics has traditionally focused on certain types of processes only, and does not provide a method for properly modeling *folding*. This paper introduces a new AI method for reasoning about the folding process. We present a discrete model for predicting a folding structure and for qualitatively simulating the process of the structure formation over time. We use a simple and approximate model for the task, but accuracy in the prediction is achieved by capturing knowledge from several sources. The model has been implemented in a working program and has been successfully tested on several types of RNAs, including RNAs whose structures have not been fully determined yet.

1 Introduction

RNA is a single-stranded molecule consisting of a sequence of bases. However, the underlying linear structure of RNA is embedded in a complex *secondary structure* because in three-dimensional space it is possible for bases in the molecule to form chemical bonds with other bases in the molecule which are not neighbors in the underlying linear structure. Determining the secondary structure ("folding") of an RNA molecule from its underlying linear structure is a complex problem involving both spatial reasoning and the use of knowledge of chemistry and biology.

Both spatial reasoning and scientific modeling have been studied before by artificial intelligence researchers. However, the problem of determining RNA secondary structure does not fit smoothly into the traditional qualitative physics framework. We believe that this domain presents challenges which require new research in both spatial reasoning and reasoning about physical systems.

RNA forms a folding structure by base pairing between complementary bases of the molecule. The structure can be determined directly by X-ray crystallogra-

phy, but most RNAs cannot be crystallized with current technique. Since the structure of RNA greatly affects its function, predicting the RNA structure is important for understanding the RNA. For an RNA of N bases, the number of different base pairs is $N(N-1)/2$. Each base pair can belong to a helix or not. Thus, the total number of possible structures is $O(2^{N^2})$. An exhaustive search for every possible structure is inadequate and impossible for this problem. We use a heuristic algorithm with folding constraint propagator for searching for possible structures. The searching process of the heuristic algorithm is guided by a scoring function whose parameters are obtained from several sources of knowledge. Since none of these sources alone provides complete and accurate knowledge for the task, generalized knowledge from the sources is incorporated into our model. As a result, we use a simple model, which is a combination of rough approximation of the original knowledge, rather than relying on any single source of knowledge as the sole determinant of the prediction. This simple, approximate model achieves much better performance in the prediction than a complex model.

2 Problem Formulation

The kinds of problems our system solves are defined as follows.

Given: an alignment of homologous RNA sequences

Find: the folding structures conserved in the sequences and the pathways from an initial structure to the final structures as the RNA sequences grow over time. Each structure should satisfy the following folding constraints [Sankoff *et al.*, 1983]:

- If the structure contains base pairing (i, j) , neither (i, k) $k \neq j$ nor (k, j) $k \neq i$ can be contained in the structure.
- If $i < j < k < l$, the structure cannot contain (i, k) and (j, l) at the same time.
- Base pairing (i, i) is not allowed for all i .
- The number of unpaired bases in a hairpin loop cannot be less than three.¹

¹A hairpin loop is one of single-stranded structure elements which folds into a shape like a hairpin. See Fig. 1D for a hairpin loop. A hairpin loop with less than 3 unpaired bases makes a too sharp turn to be stable [Tinoco *et al.*, 1973].

An RNA molecule is made up of ribonucleotides linked together in a chain. Each ribonucleotide comprises one of four different bases, adenine (A), cytosine (C), guanine (G), and uridine (U). Single-stranded RNA often folds back onto itself in structures stabilized by hydrogen bonds between A and U and between G and C. The A-U and G-C base pairs are called canonical pairs. The so-called G-U wobble pair is also frequently found although it may have a weaker bond than the canonical pairs. These two types of pairings are considered between any two bases and all the other types are considered as mismatches.

	A	C	G	U	-
A				canonical	
C			canonical		
G		canonical		wobble	
U	canonical		wobble		
-					

Table 1: Types of base pairings between two bases. Minus sign (-) indicates a gap in the sequence.

We make the following assumptions to solve the problem:

1. At least two homologous sequences are available and they are properly aligned.
2. Homologous RNA sequences have similar folding structures.

The folding structure is described in a discrete model because continuously varying parameters such as bond lengths or angles are not considered; either a hydrogen bond exists between two bases or it does not. Once the connectivity relation is determined, the shape of a secondary structure is unambiguously determined as we show in the next section.

3 Modeling the Structure Formation Process

In the RNA folding structure, each structure element contributes to the total structural stability. Helices formed by consecutive base pairings contribute to increased stability, while single stranded parts destabilize neighboring helices. Many experiments have been carried out with chemically synthesized RNAs to obtain the structural stability measured by the loss in free energy in kcal/mol [Tinoco *et al.*, 1973]. A data set of free energy values describing the contribution of each structure element is called an *energy model*. Several energy models are available but none of them can be considered perfect. Predicted structures are often radically different if a different energy model is used or a slight change is made to the energy values. Since the energy model is believed to be inaccurate as well as incomplete, we do not make commitment to the numeric values of the energy model. Instead, we use approximation of the model and seek other possible sources of knowledge, such as homologous RNAs, biochemical experimental data, and biological property of RNA. The knowledge from biochemical experiments or biological properties is not assumed to be available all the time, but it can be incorporated into

the model whenever it is available. Having such prior knowledge enables pruning the search space.

The reasoning process of our model proceeds in three phases: (1) comparative analysis of homologous sequences, (2) construction of structure model, and (3) simulation of the process of structure formation over time. Reasoning process of the model is made explicit so that the sensitivity of the model to parameter values can be examined. In this section, we describe each of the reasoning process in detail.

3.1 Comparative Analysis of Homologous Sequences

Homologous sequences are analyzed to identify potential helices, which usually maintain complementarity in the homologous sequences. Suppose that an RNA sequence contains a subsequence GAGCU. GAGCU has the potential to base pair with any AGCUC within the RNA and such a subsequence may occur several times in that RNA. In order to identify which occurrence of AGCUC actually pairs with GAGCU, the homologous RNA sequences from different organisms are examined.

```
sequence 1: GAGCU ... AGCUC ... AGCUC ... AGCUC
sequence 2: GAUCU ... AGCUC ... AGCUC ... AGAUC
sequence 3: GAUCU ... AGCUC ... AGCAC ... AGAUC
sequence 4: GACCU ... AGCUC ... AGCAC ... AGGUC
```

In the above example, only the last occurrence of AGCUC changes to maintain the ability to pair with GAGCU. Such "compensating base changes" at two positions are considered as evidence of the presence of a helix. Given an alignment of M homologous sequences of maximum length N , the program examines the sequences to identify potential base pairings based on the following algorithm:

```
for each base  $i$  in the alignment
  for each base  $j(>i)$  in the alignment
    determine the relation  $BP(i, j)$ .
```

The relation $BP(i, j)$ represents the type of base pairing between base i and base j for all the M sequences in the alignment, not for a single sequence. Since different sequences can have different types of base pairings, the type of canonical or wobble pair cannot be used for this relation. Instead, we define the relation $BP(i, j)$ as follows:

1. If base i pairs with base j in all the sequences and there is no variation in both i and j , $BP(i, j)$ is an exact-invariant pair (denoted by $\circ$).
2. If base i pairs with base j in all the sequences and there is a compensating base change in i and j , $BP(i, j)$ is an exact-variant pair (*).
3. If base i forms a G-U wobble pair with base j in most sequences, $BP(i, j)$ is a wobble pair (w).
4. If base i pairs with base j in most sequences but not in all the sequences and the frequency of the mismatches does not exceed a specified number, $BP(i, j)$ is an inexact pair (+).
5. If the frequency of mismatches between base i and base j exceeds a specified number, $BP(i, j)$ is a mismatch ($\cdot$).

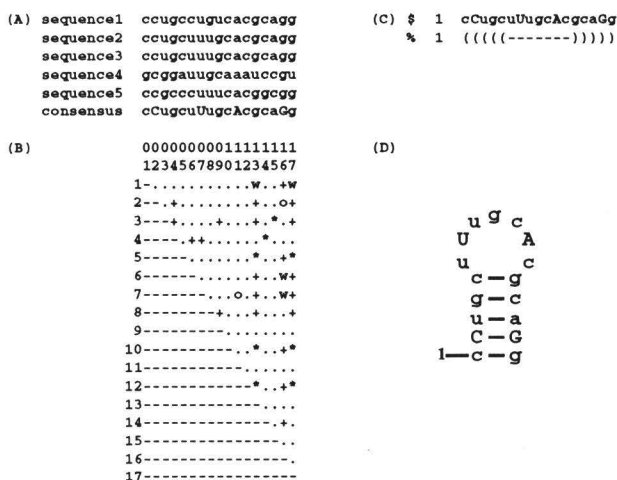


Figure 1: (A) Alignment of 5 homologous sequences and their consensus sequence. (B) Covariation matrix of the sequences. (C) Folding structure inferred from the matrix. (D) Structure prepared with a drawing program.

The result of the comparative analysis is represented in an $N \times N$ covariation matrix, where entry (i, j) represents a relation $BP(i, j)$. Since the relation $BP(i, j)$ is symmetric, only upper triangular part of the matrix is used. In the covariation matrix a potential helix of RNA appears as a diagonal of non-mismatch symbols ($\circ$, $*$, w , or $+$) from upper right to lower left. Fig. 1A shows the alignment of five RNA sequences and Fig. 1B shows the result of the comparative analysis of them. The last sequence in the alignment is a consensus sequence of the five homologous sequences, derived by the program. Upper case letter in the consensus sequence represents a base conserved in all the sequences and lower case represents a base which is not conserved in all the sequences but most representative one in the position. The consensus sequence is used for drawing the structure in this example. Although there are several non-mismatch symbols in the matrix of Fig. 1B, there is only one potential helix (1-5/13-17) which consists of at least 3 base pairings, and the folding structure inferred from the matrix is shown in Fig. 1C, represented in a form of matching parentheses. The representation is directly readable by a drawing program for a better quality representation² (Fig. 1D). Notice that any $L \times L$ submatrix ($L \leq N$) of the covariation matrix contains potential helices at a point when the RNA sequence grows to have L bases. As we show below, we use a series of submatrices to model the structures at intermediate stages during the RNA sequence growth.

3.2 Construction of Structure Model and Simulation

In the previous section, we discussed the use of a covariation matrix in finding a folding structure. In many

²The RNA structures in this paper were drawn with LoopViewer by D. G. Gilbert at Indiana University.

cases, however, a folding structure cannot be determined directly from the matrix due to conflicts among potential helices. For example, consider a helix (1-5/14-18) and a helix (1-5/19-23) in the covariation matrix of Fig. 2B. The two helices conflict each other because bases 1-5 cannot pair with both bases 14-18 and bases 19-23 at the same time (first folding constraint in Section 2). The system uses a heuristic algorithm with spatial constraint propagator incorporated in it. The algorithm first identifies all the potential helices which are not self-conflicting (i.e. the helix does not contain any base pairing (i, i) and the length of hairpin loop is at least 3). For each potential helices identified, two types of scores are computed, which are weighted sums of five parameters: helix length, number of exact-variant pairs, number of wobble pairs, number of inexact pairs, and the length of a possible hairpin loop created by the helix. S_1 is computed for finding the most stable hairpin loop-stem structure of the overall structure, and S_2 is for measuring the "goodness" of a helix in general positions (e.g. helix next to a bulge loop, internal loop, multiple loop, or hairpin loop). A helix with the highest S_1 value is called "besthelix" in the following algorithm. If there is any other helix which forms the same stem structure with the besthelix (i.e. these helices are intervened only by a bulge or an internal loop), it becomes a supporting helix of the besthelix. The besthelix and its supporting helix (if any) constitute a skeleton of the overall structure. Once a skeleton is found, it is used to eliminate the helices which violate the spatial constraints with respect to the skeleton. Each of the remaining helices is examined in decreasing order of S_2 , and is attached to currently maintained structures if it fits them. If it cannot fit, an alternative structure is created for it. The alternative structure initially contains the skeleton and the helix which had conflicts. The heuristic algorithm is briefly stated below. Propagation of the folding constraints occur in step 4 and 6 of the algorithm. The knowledge about the scoring functions comes from energy model and comparative analysis of homologous RNAs, which will be described in detail in section 4.

1. bestscore = $-\infty$
2. for each potential helix i identified
compute S_1 and S_2
if ($S_1 > \text{bestscore}$)
besthelix = helix i
bestscore = S_1 of helix i
3. skeleton = {besthelix} $\cup$ {any supporting helix}
4. Remove the helices which conflict with the skeleton
5. Sort the remaining helices with respect to S_2
6. for each of the sorted helices
if it fits current structures, add it to them
otherwise, create an alternative structure for it

Our system finds 1 - 5 common foldings of M sequences of length N in $O(MN^2 + N^3)$ time and $O(N^2)$ space. This is the same requirement as an ordinary dynamic program for folding a single sequence [Zuker and Stiegler, 1981]. For the process modeling, structures at each intermediate stage of L bases ($L \leq N$) can be found in

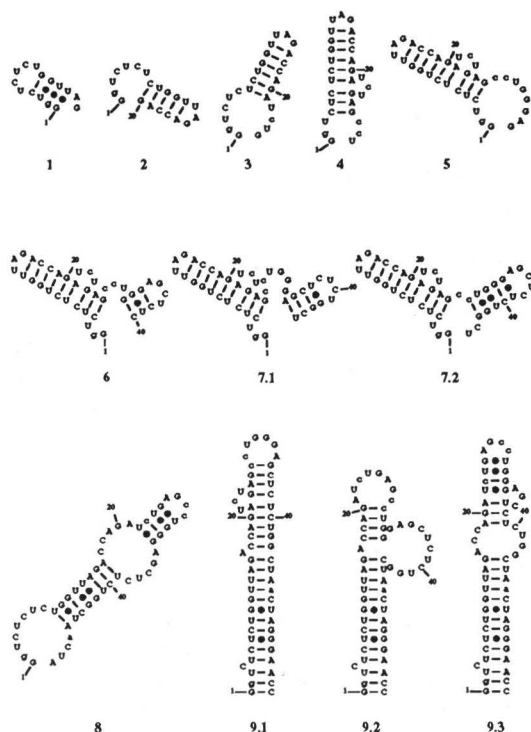


Figure 4: Structure of TAR at intermediate stages

structures in each stage instead of the ones which have the easiest transition from the previous stage.

To obtain a stable common structure, it is generally better to look at more sequences. However, if there is little variation between the sequences, a large number of sequences are not necessary. The value of minimum length of a helix greatly affects the number of potential helices. The smaller the value, the more the potential helices. It also influences the prediction. If the number of minimum length of a helix is set to a larger value than 3, then a rough structure is obtained where helices shorter than the threshold value are missing but all the other helices stay. Therefore, increasing the minimum length of a helix does not affect much the prediction other than that shorter helices are ignored. Decreasing the number of minimum length of a helix from 3 to 2 has a slightly more complicated influence on the prediction; it can simply add new helices of length 2 to the previous structure while preserving all the existing helices in it; helices of length 2 can sometimes, however, defeat longer helices in the previous structure and replace them. The value of number of mismatches allowed is another parameter which controls the number of potential helices. The larger the value, the more the potential helices. In most cases we tested, allowing mismatches within 10% (e.g. 1 mismatch for 10 sequences) produces satisfactory results. However, the optimum value of the mismatches allowed will depend on how much variation exists between the sequences and how representative they are.

Weight values of the scoring function do not affect the number of potential helices at all, but has a direct and most influential effect on the prediction. Longer helices

are more stable than shorter ones and exact-variant pairs are stronger evidence to base pairing than other types of pairs. Hence, the length and frequency of exact-variant pairs are assigned positive weights in the scoring function. Wobble pairs and inexact pairs are less stable than canonical pairs, and longer hairpin loops are less stable than shorter ones. Therefore, they have negative weights. While decision on the signs of the five parameters is straightforward, decision on the absolute values of them is not. Currently we do not have a general theoretic relation which explains how the relative sizes of the five values are correlated with respect to the prediction. The parameter values in Fig. 3 have been obtained experimentally. We have successfully tested the program on different types of RNAs (tRNAs, 5S rRNAs, 16S rRNAs, TAR, RRE, and *cis*-acting packaging sequences of HIV-1), in many of which a widely used conventional program fails. A detailed comparison is described in [Han and Kim, 1993]. The execution time for predicting common foldings of 10 tRNAs of length 77, 34 5S RNAs of length 143, 10 16S rRNAs of length 1651, 10 TAR RNAs of length 57, 10 RRE RNAs of length 196, and 10 HIV-1 *cis*-packaging sequences of length 113 are 0.1, 0.6, 5.1, 0.1, 0.4, and 0.2 seconds, respectively.

5 Related Work

There currently exists no reliable and automatic method for modeling the folding process and for predicting folded structures. Existing approaches have limitations in their use either because they are computationally too expensive to be practical, their predicted structures are not stable or they heavily rely on manual work [Brown, 1991; Noller and Woese, 1981; Zuker and Stiegler, 1981]. The algorithm by Sankoff[1985] for simultaneously aligning and folding M sequences of length N requires $O(N^{3M})$ time and $O(N^{2M})$ space, and is not practical either in terms of time or space.

Like us, several researchers in AI have used approximation in reasoning. Weld[1986] describes a simulator which uses aggregation as an abstraction technique for a continuous process. His approach cannot be applied to simulate the folding process because the process depends on the constituent bases rather than on any regularity of the process, such as repeating cycles. Most research in qualitative kinematics has focused on predicting systems' behavior from the description of the systems [Faltings, 1987; Forbus *et al.*, 1991; Gelsey, 1991; Joskowicz, 1989]. A general constraint common in those systems is that objects cannot occupy the same space. The RNA folding problem has a similar constraint in the sense that a base cannot pair with more than one base. However, the current technique of qualitative kinematics seems inapplicable to the RNA folding problem.

6 Future Work

The model presented in this paper has limitations and offers several directions for future research: (1) The process of examining the effects of parameter changes of the model on the result is not automated. We plan to develop methods, both theoretical and experimental, for

automating this process. This will enable us to do more extensive perturbation study on the parameters affecting the behavior of the model. (2) We currently use a discrete model both for the structure and for the structure formation process. For the structure, a discrete model is more appropriate than a continuous model because there are no continuously varying parameters to consider in the model by its nature. The process of the structure formation is also modeled in a discrete model by taking "snapshots" of the structures at a series of intermediate stages. We might want to consider a continuous model for the process, which enables us to take into account the rate that a structure unwinds its helices to form a new helix as RNA grows. (3) The current model considers secondary interactions only and predicts a secondary structure. The model can be extended to incorporate tertiary interactions such as pseudo knots, which are base pairings between a loop and another part. Some of conflicting diagonals in the covariation matrix constitute pseudo knots. (4) The current model can be easily applied to find disulfide bonds of protein structure as well. The model should be extensible to other general problems which involve matching or compatibility between objects. Any relation of compatibility or conflict between objects is represented in a covariation matrix, and a searching algorithm operates on the matrix to find all the possible solutions which satisfy the relation. The extended model is currently being implemented.

7 Conclusion

In this paper, we have described a new qualitative model for the folding problem of a biological system, and demonstrated that the model can be used both for predicting structures and for simulating the process of the structure formation over time. This model is a rough approximation of conventional models, but greater accuracy in the prediction is obtained by using knowledge from diverse sources instead of relying on a single source of knowledge, which is often inaccurate and incomplete. Homologous RNA sequences are analyzed simultaneously for filtering potential helices and the result is incorporated with knowledge of free energy. A heuristic algorithm with folding constraint propagator is used for searching for possible structures. The most plausible structures conserved in multiple homologous sequences are produced in the same amount of time and space as those of the conventional models for a single sequence. The model has been successfully tested on various types of RNAs. It is known that there is no reliable or automatic way of predicting a common folding even if the related sequences are already aligned and that one way to obtain a common folding is to fold each sequence separately using an ordinary folding program and to search for common structures [Zuker, 1989]. Our model is the first one provably practical both in terms of time and space for finding folding structures of homologous RNAs.

Acknowledgments

We thank Mick Noordewier and Hong-Jin Kim for their contributions to this effort.

References

- [Brown, 1991] James W. Brown. Phylogenetic comparative analysis of RNA structure on Macintosh computers. *CABIOS*, 7(3):391–393, 1991.
- [Cullen, 1990] Bryan R. Cullen. The HIV-1 Tat Protein: An RNA Sequence-Specific Processivity Factor? *Cell*, 63(16):655–657, 1990.
- [Faltings, 1987] Boi Faltings. Qualitative Kinematics in Mechanisms. In *Proc. of IJCAI-87*, pages 436–442, 1987.
- [Forbus *et al.*, 1991] Kenneth D. Forbus, Paul Nielsen, and Boi Faltings. Qualitative spatial reasoning: the CLOCK project. *Artificial Intelligence*, 51:417–471, 1991.
- [Gelsey, 1991] Andrew Gelsey. The Use of Intelligently Controlled Simulation to Predict a Machine's Long-Term Behavior. In *Proc. AAAI-91*, pages 880–887, 1991.
- [Han and Kim, 1993] Kyungsook Han and Hong-Jin Kim. Prediction of Common Folding Structures of Homologous RNAs. *Nucleic Acids Research*, 21(5):1251–1257, 1993.
- [Joskowicz, 1989] Leo Joskowicz. Simplification and Abstraction of Kinematic Behaviors. In *Proc. IJCAI-89*, pages 1337–1342, 1989.
- [Noller and Woese, 1981] Harry F. Noller and Carl R. Woese. Secondary Structure of 16S Ribosomal RNA. *Science*, 212(4493):403–411, 1981.
- [Sankoff *et al.*, 1983] David Sankoff, Joseph B. Kruskal, Sylvie Mainville, and Robert J. Cedergren. Fast Algorithms to Determine RNA Secondary Structures Containing Multiple Loops. In *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, pages 93–120. Addison-Wesley Publishing Company, 1983.
- [Sankoff, 1985] David Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM journal on applied mathematics*, 45(5):810–825, 1985.
- [Tinoco *et al.*, 1973] Ignacio Tinoco, Jr., Philip N. Borer, Barbara Dengler, Mark D. Levine, Olke C. Uhlenbeck, Donald M. Crothers, and Jay Gralla. Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nature New Biology*, 246:40–41, 1973.
- [Weld, 1986] Daniel S. Weld. The Use of Aggregation in Causal Simulation. *Artificial Intelligence*, 30:1–17, 1986.
- [Zuker and Stiegler, 1981] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.
- [Zuker, 1989] Michael Zuker. Computer Prediction of RNA Structure. In James E. Dahlberg and John N. Abelson, editors, *Methods in Enzymology*, volume 180, pages 262–288. Academic Press, Inc., 1989.