

Computer RNA Three-Dimensional Modeling From Low Resolution Data and Multiple-Sequence Information

François Major, Sébastien Lemieux and Abdelmjid Ftouhi

Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal
Montréal, Québec, Canada H3C 3J7
major@iro.umontreal.ca

The problem of modeling three-dimensional structures of ribonucleic acids is expressed in terms of the constraint satisfaction problem. Three-dimensional structures are represented by constraint graphs, where vertices represent nucleotides and edges represent structural constraints. A formalism to help rationalize a series of modeling experiments in the context of low resolution and multiple-sequence data was developed. From secondary structure and low resolution data, several structural hypotheses corresponding to different constraint graphs can be derived. In presence of several structurally related sequences, the application of three-dimensional modeling to each sequence and hypotheses produces a sequence-structure relation that can be analyzed using fuzzy set theory, given the imprecision and uncertainty involved in the modeling process.

The popularity of computer modeling of RNA three-dimensional structure can be explained by the desire to rapidly understand the function of newly discovered RNAs and by the difficulties of applying high resolution structure determination techniques, such as X-ray crystallography and nuclear magnetic resonance spectroscopy. Computer modeling implies the interpretation of experimental data, the formation of structural hypotheses, and the building of three-dimensional models. Such models offer a simultaneous view of many aspects of the molecule and allow one to design more precise and incisive experiments which, in turn, generate new structural data and hypotheses leading to new modeling experiments. Thus, models are dynamic objects that represent the quantity and quality of structural knowledge on a molecule at a given time. The iterative use of modeling and low resolution experimental methods should converge on a highly defined and accurate model.

FM is a fellow of the Canadian Genome Analysis and Technology program and the MRC of Canada. This work has been supported by the MRC of Canada.

Most three-dimensional modeling projects begin with primary and secondary structure, low resolution data and multiple-sequence data (I) from which many different structural hypotheses can be derived. One way to support a structural hypothesis consists in building a consistent three-dimensional model compatible with each available sequence (2, 3, 4). A systematic verification consists in building all possible models for each active sequence. This creates a relation, $R \subseteq S \times H$, where R , the set of relations, is a subset of binary relations between S , the set of sequences, and H , the set of structural hypotheses; $(S_i, H_j) \in R$ if and only if the sequence S_i generates at least one three-dimensional model that satisfies the structural hypothesis H_j . Each structural hypothesis, H_j , is associated with a set, E_{H_j} , that contains all three-dimensional models consistent with H_j . Computer programs such as MC-SYM (5) which transform constraint graphs into three-dimensional models can be used, although the formalism presented here is independent of any particular modeling method.

Uncertainty in modeling lies in the fact that a three-dimensional model can either support a structural hypothesis or can be the result of modeling artifacts. Computer modeling is subject to imprecision in the low resolution data, subjectivity in the generation of three-dimensional models, and uncertainty in the formation of structural hypotheses. The theory of possibility, based on fuzzy logic, is used to classify structural hypotheses according to their likelihood to contain multiple-sequence data consistent conformations based upon the sequence-structure relation, R .

In this article we present the constraint graph representation used by MC-SYM to transform structural data into three-dimensional models. Then, we discuss the sequence-structure relation and the theory of possibility to assign plausibility coefficients to each structural hypothesis. Finally, we discuss the application of this technique to the lead-activated ribozyme and indicate how modeling was used iteratively with experimentation to derive its active structure.

RNA Conformational Space

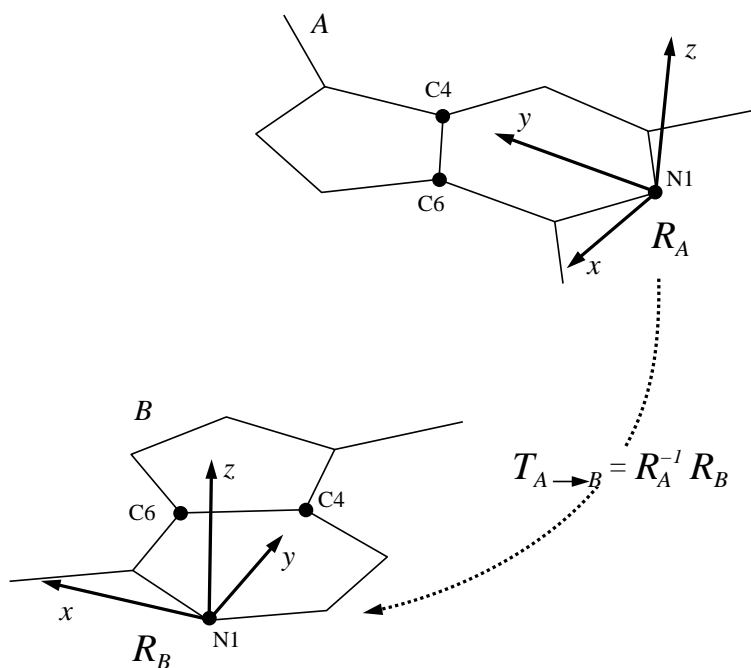
Here, we consider a *RNA three-dimensional structure* as the assembly of its constituent nucleotides in three-dimensional space. We introduce a *RNA conformational search space* defined by molecular contacts (or constraints). The molecular contacts are used in operators that position and orient the nucleotides in three-dimensional space.

A *molecular contact* is formed between two nucleotides, A and B , if they are connected through a phosphodiester bond or if they share a hydrogen bond between their nitrogen bases. The combination of all molecular contacts constitutes the *contact graph* of the RNA. It is self evident from the definition of a molecular contact that in all known RNA three-dimensional structures, every nucleotides make at least one molecular contact with another one. Thus, all RNAs contain at least one *path* of molecular contacts that connects all its constituent nucleotides, which does not contain any cycle, a *spanning tree* of the *nucleotide contact graph*.

In the following, we first present how contact graphs define the conformational search space of RNAs, to position and orient all nucleotides in three-dimensions. Then, a database of spatial relations based on molecular contacts, as observed among pairs of nucleotides in known structures, is introduced.

RNA Conformational Search Space Defined by Molecular Contacts. The premise to use molecular contacts in defining the conformational space of RNAs relies on the fact that molecular contacts contain all the information critical to the global fold of the molecule. Consider the best characterized case of an RNA double-helix. The spatial relation between two bases involved in a Watson-Crick pairing can be used, in conjunction with a canonical base stacking geometry, as a good approximation to position and orient double-helical strands in three-dimensions.

The spatial information is encoded by homogeneous transformation matrices (6). The *local referential* of a nucleotide, A , can be represented by an homogeneous transformation matrix, R_A . R_A is determined by the coordinates of three atoms in A from which three right handed unary orthogonal vectors can be derived. The Cartesian coordinates of the first selected atom, for instance, can be chosen as the origin of the residue (see Figure 1). The *spatial relation* between two nucleotides, A and B , is an homogeneous coordinate transformation matrix, $T_{A \rightarrow B} = R_A^{-1} R_B$. In this way, the spatial relations between any pair of nucleotides forming molecular contacts in the known three-dimensional structures can be extracted and used as building blocks of RNA three-dimensional structures.



An observed contact between nucleotides A and B can be reproduced between any pair of nucleotides, let's say A' and B' , by applying the homogeneous transformation matrix $R_{B'}^{-1} T_{A \rightarrow B} R_{A'}$ to the atomic coordinates of B' to position and orient B' with respect to A' as observed between nucleotides A and B ; or symmetrically, by applying the homogeneous transformation matrix $R_{A'}^{-1} T_{A \rightarrow B}^{-1} R_{B'}$ to the atomic coordinates of A' . The final result of this manipulation is, in either case, that the new observed spatial relation between A' and B' is exactly the same as that observed between A and B , thus reproducing the same molecular contact in the newly built model.

A *transformational set* is a set of homogeneous transformation matrices associated with a molec-

ular contact type defined by the nature of the nucleotides in contact. For instance, there are four types of RNA bases determining ten different types of pairs by considering that symmetric pairs are likely to share the same types of molecular contacts, and two main types of molecular contacts: paired and connected. Connected nucleotides can be either stacked or not. In practice, we consider only two types of bases (purines and pyrimidines) for contact defined by a phosphodiester bond. This partition of the different molecular contacts gives a possibility of $(10 + 6 = 16)$ different transformational sets.

The number of spanning trees, pairs of nucleotide contacts and homogeneous transformation matrices associated with a contact graph determine the conformational search space *size* of a RNA. The number of homogeneous transformation matrices associated with a molecular contact type is given by the number of occurrences observed in all available RNA three-dimensional structures in the Protein DataBank (PDB) (7), Nucleic acids DataBase (NDB) (8) and other personally communicated structures.

The transformations were extracted and classified among the 16 different types of contacts. Those sets were then sorted in such a way that any subset composed of the first n elements represents the most efficient sampling of the addressed space. This property is achieved by selecting, as the first element of the set, the one that minimizes the sum of its distances with all other elements. This element is then considered the most “common” example. The next elements are those that maximizes their distances with all previously included elements. This sorting method supposes the existence of a distance metric to evaluate the difference between two homogeneous transformation matrices. The simplest metric is to sum the squares of the differences between the corresponding matrix elements, the Euclidean distance metric.

Starting from the contact graph, an efficient way to build three-dimensional models is to first determine a reference nucleotide that will be placed arbitrarily in three-dimensional space. From that, a spanning tree of the contact graph is expanded, determining which contacts will be used in the building procedure. For each molecular contact appearing in the spanning tree, the corresponding transformational set is used to systematically search the conformational space for valid three-dimensional models. Molecular contacts represented by edges that are not considered in the selected spanning tree are replaced in the simulation by geometrical constraints to guarantee their satisfaction in the final three-dimensional models.

The computer program MC-SYM is currently used to perform this search. Since the number of spanning trees of a fully connected graph composed of N vertices is N^{N-2} , the problem of selecting the one that is the most likely to generate complete structures is still open. The fuzzy logic approach presented here was developed in part to deal with the inaccuracy introduced by the approximation made while selecting a specific spanning tree.

The Sequence-Structure Relation. Structural hypotheses are derived from available structural data and are distinguished by their patterns of base pairing and stacking. For each structural hypothesis and sequence, a three-dimensional modeling simulation is performed, for instance using the MC-SYM program. In fact, any three-dimensional scheme determining if a three-dimensional model can be built for a given sequence and constraint graph is acceptable. The sequence-structure relation is established by associating the sequences to their consistent structural hypotheses; a link is created if and only if a three-dimensional model can be built. An MC-SYM input script describes the constraint graph and a sequence. By changing the latter, one can easily verify if a different se-

quence is compatible with the constraint graph.

Terminology and Notation of the Uncertainty Principle

In the history of mathematics, *uncertainty* was approached in the XVIIth century by Pascal and Fermat who introduced the notion of probability. However, probabilities do not allow one to process subjective beliefs nor imprecise or vague knowledge, such as in computer modeling of three-dimensional structure. Subjectivity and imprecision were only considered from 1965, when Zadeh, known for his work in systems theory, introduced the notion of *fuzzy set*. The concept of fuzziness introduces partial membership to classes, admitting intermediary situations between no and full membership. Zadeh's *theory of possibility*, introduced in 1977, constitutes a framework allowing for the representation of such *uncertain* concepts of non-probabilistic nature (9). The concept of fuzzy set allows one to consider imprecision and uncertainty in a single formalism and to quantitatively measure the preference of one hypothesis versus another. Note, however, that Bayesian probabilities could have been used instead.

Consider a finite reference set, X . Events can be defined by subsets of X to which can be assigned coefficients between 0 and 1 evaluating their possibility to occur. In order to define these coefficients, a measure of possibility is introduced, Π , which is a function defined over the power set of X (the set of all subsets composed of the elements of X), $\mathcal{P}(X)$, the parts of X which take their values in $[0, 1]$, such that:

$$\Pi(\emptyset) = 0, \quad \Pi(X) = 1, \quad (1)$$

$$\begin{aligned} \forall A_1 \in \mathcal{P}(X), \quad A_2 \in \mathcal{P}(X), \dots \\ \Pi(\bigcup_{i=1,2,\dots} A_i) = \max_{i=1,2,\dots} \Pi(A_i), \end{aligned} \quad (2)$$

where \emptyset is the empty set and \max indicates the maximum value of all values. The possibility associated to the empty set is zero. The possibility of X is one. The possibility of a series of events (union) is the maximum possibility among these events.

The functions of *belief* concern a quantification of credibility attached to the events. Shafer's *theory of evidence* considers a finite universe of reference, X , upon which are determined belief coefficients obtained by distributing a global mass of belief equal to 1 among all possible events (10). A mass, m , can be defined as follows:

$$m : \mathcal{P}(X) \longrightarrow [0, 1]$$

such that

$$m(\emptyset) = 0 \quad \text{and} \quad \sum_{A \in \mathcal{P}(X)} m(A) = 1.$$

For each set $A \in \mathcal{P}(X)$, the value $m(A)$ represents the degree with which a group of observers believe in the realization of an event from the elements of A . This value, $m(A)$, involves only a single set, the set A , and does not involve any other information for the subsets of A . If there exists additional evidence which confirms the realization of the same event in a subset of A , $B \subset A$, it must be expressed by another value, $m(B)$. Every non empty part A of X , for which $m(A) \neq 0$,

is called a *focal element* corresponding to an event believed by the observers. The *belief measure* of such a part A of X is defined by considering all the focal elements implying A :

$$Bel(A) = \sum_{B|B \subseteq A} m(B),$$

that is, the belief of a part A of X is defined by the sum of all parts, B such that A contains B . The *plausibility measure* of A is defined by taking all focal elements related to A :

$$Pl(A) = \sum_{B|B \cap A \neq \emptyset} m(B),$$

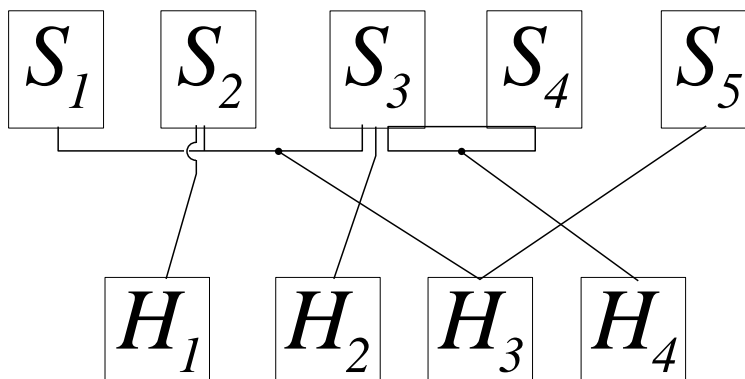
that is, the plausibility of a part A of X is defined by the sum of all parts, B such that B intersects with, or contains any element of, A . The above measures verifies the following relations:

$$Pl(A) = 1 - Bel(\bar{A}) \quad \text{and} \quad Bel(A) \leq Pl(A).$$

where \bar{A} indicates the complement of A in X .

The range $[Bel(A), Pl(A)]$ embeds the imprecise probability, $P(A)$, for any part A of X . A particular case of the mass m is remarkable: consider that all focal elements are singletons of X , that is, beliefs only concern elementary events. Then, every part A of X is such that $Bel(A) = Pl(A)$ and this common value is equal to the probability, $P(A)$.

Calculating Possibilities. Consider the sequence-structure relation in Figure 2. A uniform probability of $\frac{1}{5}$ is assigned to each sequence. Note that the uniform distribution is not a requirement of the mathematical model. It was assumed that all sequences could adopt the same conformation. The possibility for each structural hypothesis to contain the conformation was computed using Zadeh's theory of possibility.



Consider X as the set containing all conformations generated by MC-SYM for all sequence variants. From the sequence-structure relation, R , the focal elements are

$$S_{1,5}, S_2, S_3 \text{ and } S_4,$$

where $S_{1,5} = S_1 \cap S_5$ and S_1, S_2, S_3, S_4, S_5 are subsets of X which contain the three-dimensional conformations associated with the active structure.

From R we have:

$$\begin{aligned} S_{1,5} &= E_{H_3} \\ S_2 &= E_{H_1} \cup E_{H_3} \\ S_3 &= E_{H_2} \cup E_{H_3} \cup E_{H_4} \\ S_4 &= E_{H_4} \end{aligned}$$

where $E_{H_i}, i = 1, 2, 3, 4$, represents the set of conformations that satisfy hypothesis H_i .

A belief coefficient of possibility to contain the conformation is assigned to each focal element:

$$\begin{aligned} m(S_{1,5}) &= \frac{2}{5} \\ m(S_2) &= m(S_3) = m(S_4) = \frac{1}{5}. \end{aligned}$$

The basic probabilities (masses) were assigned by considering that any of the available sequences could adopt the active conformation. The possibility distribution, π , is then

$$\forall x \in X \quad \pi(x) = 1,$$

which is equivalent in the case of the probability distribution, p , to

$$\forall x \in X \quad p(x) = \frac{1}{|X|}.$$

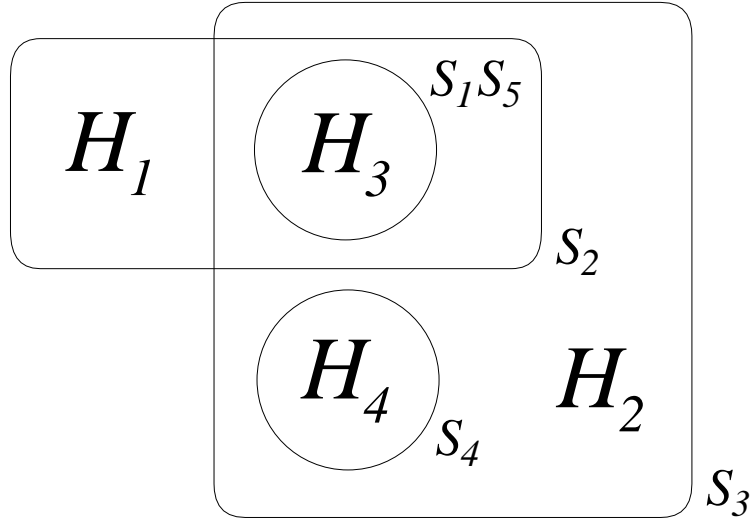
However, given the biological supposition that the active conformation should be found among the structures common to all sequences, the belief coefficients were assigned according to how many sequences are compatible with the hypothesis, that is, for $S_{1,5}$, the belief coefficient of sequences S_1 , and S_5 ,

$$m(S_{1,5}) = \frac{2}{5},$$

based on the fact that two sequences in the set of five sequences were found compatible with a particular subset of the structural hypothesis. Figure 3 shows the focal elements $S_{1,5}, S_2, S_3$ and S_4 . It is now possible to define the intervals of probabilities (possibilities) for each part of the conformational space.

This situation allows to deduce the intervals of probabilities of each structural hypothesis, that is, the belief measures:

$$\begin{aligned} Bel(E_{H_1}) &= Bel(E_{H_2}) = 0 \\ Bel(E_{H_3}) &= \sum_{B|B \subseteq E_{H_3}} m(B) = m(S_{1,5}) = \frac{2}{5} \\ Bel(E_{H_4}) &= \sum_{B|B \subseteq E_{H_4}} m(B) = m(S_4) = \frac{1}{5} \end{aligned}$$



and the plausibility measures:

$$\begin{aligned}
 Pl(E_{H_1}) &= \sum_{B|B \cap E_{H_1} \neq \emptyset} m(B) \\
 &= m(S_2) = \frac{1}{5} \\
 Pl(E_{H_2}) &= \sum_{B|B \cap E_{H_2} \neq \emptyset} m(B) \\
 &= m(S_3) = \frac{1}{5} \\
 Pl(E_{H_3}) &= \sum_{B|B \cap E_{H_3} \neq \emptyset} m(B) \\
 &= m(S_{1,5}) + m(S_2) + m(S_3) = \frac{4}{5} \\
 Pl(E_{H_4}) &= \sum_{B|B \cap E_{H_4} \neq \emptyset} m(B) \\
 &= m(S_3) + m(S_4) = \frac{2}{5}
 \end{aligned}$$

which are summarized in Table I. According to the belief and plausibility coefficients calculated for all the hypotheses, H_3 , with an imprecise probability over the range $[Bel(E_{H_3}), Pl(E_{H_3})]$, is the one that seems the most likely to be shared by all variant sequences. This does not necessarily indicate, without any doubt, that the actual three-dimensional structure will be found in those of H_3 . It simply indicates that among the current evaluated hypotheses H_3 is the one that best reflects the combined results of the modeling experiments. With that information in hand, the models generated under H_3 should be carefully examined and used in the design of future laboratory experiments, either to confirm the hypothesis or produce new structural data and hypotheses.

Table I. Belief and Plausibility Measures for H_1 , H_2 , H_3 and H_5 . $Bel(A)$ is the belief value. $Pl(A)$ is the plausibility value.

A	$Bel(A)$	$Pl(A)$
E_{H1}	0	$\frac{1}{5}$
E_{H2}	0	$\frac{1}{5}$
E_{H3}	$\frac{2}{5}$	$\frac{4}{5}$
E_{H5}	$\frac{1}{5}$	$\frac{2}{5}$

Application to the Leadzyme

The Pb^{2+} cleavage of a specific ribophosphodiester bond in yeast tRNA^{Phe} is the classical model of metal-assisted RNA catalysis. *In vitro* selection experiments have identified tRNA^{Phe} variants a derivative of which, named the leadzyme, is very active in cleavage by Pb^{2+} (11). The leadzyme consists of an RNA duplex with an asymmetric internal loop of six nucleotides (12). Cleavage of the leadzyme domain produces two fragments: one with a terminal 5'-hydroxyl group, and the other with a 3'-phosphomonoester presumably generated via a 2',3'-cyclic phosphodiester intermediate. The two-step reaction mechanism of the leadzyme is reminiscent of protein ribonucleases and distinguishes it from other ribozymes, such as the hammerhead, the hairpin and the hepatitis δ domains, which produce 2'-3'-cyclic phosphates (13, 14, 15, 16). The detailed three-dimensional structure was a requisite to the understanding of the particularities of this reaction.

Modeling of the leadzyme was initiated with a series of structural hypotheses derived from the primary and secondary structures. A list of active analogous sequences was previously isolated by *in vitro* selection experiments (11). The program MC-SYM was used to establish the sequence-structure relation, $R \subseteq S \times H$ by generating conformational libraries for the wild-type sequence and all sequence analogs. The fuzzy logic model was applied to these libraries to identify the most plausible hypothesis that was then experimentally evaluated. Activity data of leadzyme variants that incorporated modified nucleotides into the catalytic core (17) led to a new structural hypothesis and a second round of computer modeling. The final model is consistent with all available structural data and provided insight into the catalytic reaction of this ribozyme. The details about the active conformation and the three-dimensional modeling of the leadzyme are reported in a manuscript in preparation, available from the authors.

Conclusion

A mathematical model based on fuzzy logic was developed for the selection of structural hypotheses that are more likely to contain active conformations consistent with a series of analogous sequences. The application of this model is especially useful when multiple-sequence data are available, that however do not reveal sufficient structural aspects to initiate three-dimensional modeling. The MC-SYM program or any other RNA modeling approach can be used to produce the sequence-structure relation. The fuzzy logic model was incorporated in the iteration of computer modeling, hypothesis formation and experimental work. This protocol was successfully applied to the three-dimensional modeling of the leadzyme. The fuzzy logic model made possible the identification of a structural hypothesis that was used in the design of laboratory experiments which, in turn, generated structural

data that produced a final consistent model.

Literature Cited

1. Major, F.; Gautheret, D. In *Encyclopedia of Molecular Biology and Molecular Medicine*; Myers, R.A., Ed.; VCH Publishers Inc.: NY, 1996, Vol. 5; pp 371–388.
2. Brown, J.; Nolan, J.; Haas, E.; Rubio, M.; Major, F.; Pace, N. *Proc. Natl. Acad. Sci.* **1996**, 93, pp. 3001–3006.
3. Gautheret, D.; Koonings, D.; Gutell, R. *J. Mol. Biol.* **1994**, 242, pp. 1–8.
4. Michel, F.; Westhof, E. *J. Mol. Biol.* **1990**, 216, pp. 585–610.
5. Major, F.; Turcotte, M.; Gautheret, D.; Lapalme, G.; Fillion, E.; Cedergren, R. *Science* **1991**, 253, pp. 1255–1260.
6. Paul, R. P. *Robot Manipulators: Mathematics, Programming, and Control*; MIT Press: Cambridge, MA, 1981.
7. Bernstein, F. C.; Koetzle, T. F.; Williams, G.J. B.; Meyer, E.F. J.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Eur. J. Biochem.* **1977**, 80, pp. 319–324.
8. Berman, H.; Olson, W.; Beveridge, D.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A.; Schneider, B. *Biophys. J.* **1992**, 63, pp. 751–759.
9. Zadeh, L. In *Fuzzy sets and systems I*; North-Holland Publishing Company: Amsterdam, Holland, 1977, pp. 3–28.
10. Shafer, G. *A mathematical theory of evidence*; Princeton Univ. Press: Princeton, NJ, 1976.
11. Pan, T.; Uhlenbeck, O. *Biochemistry* **1992**, 31, pp. 3887–3895.
12. Pan, T.; Uhlenbeck, O. *Nature* **1992**, 358, pp. 560–563.
13. Buzayan, J.; Gerlach, W.; Bruening, G. *Proc. Natl. Acad. Sci.* **1986**, 83, pp. 8859–8862.
14. Hutchins, C.; Rathjen, P.; Forster, A.; Symons, R. *Nucl. Acids Res.* **1986**, 14, pp. 3627–3640.
15. Forster, A. C.; Symons, R. H. *Cell* **1987**, 49, pp. 211–220.
16. Epstein, L.; Gall, J. *Cell* **1987**, 48, pp. 535–543.
17. Chartrand, P.; Usman, N.; Cedergren, R. *Biochemistry* **1997**, 36, pp. 3145–3150.