Symbolic Generation and Clustering of RNA 3-D Motifs

Marielle Foucrault[1] and François Major[1,2]
[1]Département d'Informatique et de Recherche Opérationnelle
[2]Montreal Joint Center for Structural Biology
Université de Montréal
Montréal, Québec, Canada H3C 3J7
major@iro.umontreal.ca

Abstract

Non canonical G.A base pairs play important structural and functional roles in ribonucleic acids (RNA). In particular, the $\begin{matrix}5'-G-A-3'\\3'-A-G-5'\end{matrix}$ motif and three of its sequence variants have a relatively high occurrence in 16S and 23S ribosomal RNA. Extensive 3-D modeling of these variants has allowed to support a previously proposed 3-D model and to identify another series of conformations consistent with phylogenetic data. The library of 3-D conformations generated by the MC-SYM program was then used to produce 3-D conformations of the small ribonucleotide r(GGCGAGCC)$_2$. This new library includes the conformation determined by nuclear magnetic resonance spectroscopy.

Introduction

Case-based reasoning methods have been among the most successful approaches to three-dimensional (3-D) modeling of proteins mainly because of the great similarity between the 3-D structures of two homologous peptide sequences (Hilbert, Bohm, & Jaenicke 1993). In the case of ribonucleic acids (RNA), however, the small number of experimentally determined 3-D structures limits the application of this approach and alternative methods are sought (Malhotra, Tan, & Harvey 1990; Major *et al.* 1991; Altman 1993; Altman, Weiser, & Noller 1994).

RNA molecules are composed of double-stranded motifs, called double-helices, which are built of Watson-Crick base pairs and adopt a canonical A-RNA conformation. The double-helices are inter-connected by single-stranded loops and bulges (see Figure 1) which determine their position and orientation in 3-D space. Bulges (Figure 1a) are unilateral insertions of unpaired nucleotides in a double helical strand. Internal loops (Figure 1b) are single-stranded regions that connect two distinct double-helices. Hairpin loops (Figure 1c) are single-stranded RNA connecting the tips of a terminal base paired region. The observation of only few examples of RNA internal and hairpin loops by X-ray crystallography and nuclear magnetic resonance (nmr) spectroscopy suggests that their 3-D structures are variable and highly depend on their length, flanking base pairs and sequence. Non canonical base pairs of various geometrical patterns have been observed in all cases.

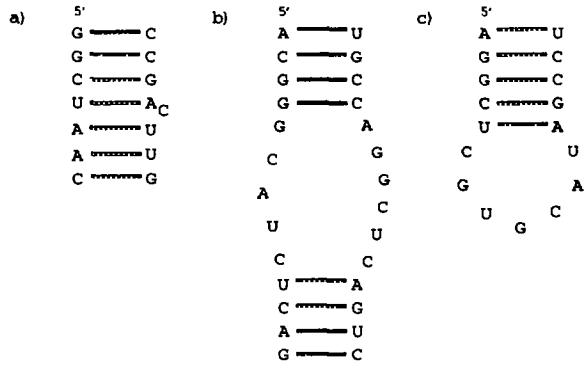

Figure 1: Bulges and loops. a) RNA double-helix interrupted by a bulge of length 1. b) Two RNA double-helices connected by an internal loop with five nucleotides in the 5' strand and six nucleotides in the 3' strand. c) RNA seven-membered hairpin loop.

Two frequently occuring hairpin loops in ribosomal RNAs (rRNA) are: GNRA (where N represents any nucleotide and R represents purines) and UUCG. Both loop structures have been studied in solution by nmr spectroscopy. The GNRA contains a non canonical G.A base pair (Heus & Pardi 1991) and the UUCG contains a non canonical G.U base pair (Cheong, Varani, & Tinoco 1990). Nmr spectroscopy and X-ray crystallography studies have also revealed the presence of non canonical base pairs in internal loops. The solution structure of loop E of Eukaryotic 5S rRNA contains a closing G.A base pair and two non canonical, Hoogsteen and reverse-Hoogsteen, A.U base pairs (Wimberly, Varani, & Tinoco 1993). The solution structure of the internal loop of HIV-1 rev-binding element contains purine-purine base pairs

This work was supported by the MRC of Canada.

(Battiste *et al.* 1994). Comparative sequence analysis of rRNA have also revealed frequently occuring tandem G.A mismatches (Gautheret, Koonings, & Gutell 1994). The tandem G.A is also present in the hammerhead (Pley, Flaherty, & McKay 1994) and lead-activated (Pan & Uhlenbeck 1992) ribozymes. In group I introns and the rev-binding element of HIV-1, non canonical base pairs have been deduced from co-variation analysis and have been successfully introduced in RNA 3-D modeling (Michel & Westhof 1990; Leclerc, Cedergren, & Ellington 1994).

Here, we report on the application of the computer modeling system MC-SYM (Major *et al.* 1991) to generate a library of tandem G.A mismatches. Such libraries are necessary prior to further structural studies such as those based on energy stability, isosteric conformations and chemical data. As a first step, the constraint satisfaction algorithm was applied to generate all possible tandem G.A mismatches conformations and their interchangeable motifs in 16S and 23S rRNAs (Gautheret, Koonings, & Gutell 1994). Large numbers of conformations were generated mainly due to the structural flexibility of non canonical base pairing and stacking of the purines. The symbolic information generated by the program was then used to classify the conformations into structural classes based on base pair geometries, sugar pucker modes and torsion angle around the glycosyl bond. The library was used to identify two series of isosteric motifs that are consistent with phylogenetic data. The first one falls in the same class as the solution structure of r(GGCGAGCC)$_2$ determined by SantaLucia et al. (SantaLucia & Turner 1993). The other, which we propose as a second possibility, falls in the same class as the solution structure determined by Privé et al. (Privé *et al.* 1987) for d(CCAAGATTGG)$_2$.

Complexity analysis of MC-SYM

The backtracking algorithm used in MC-SYM, although efficient, is based on a combinatorial search. This implies that the conformational search must be meticulously applied to avoid the combinatoric explosion. In the current implementation, the free variables are the nucleotide units. The domain of values for each nucleotide is defined by the Cartesian product of various pre-determined nucleotide conformations, that is, rigid sets of atomic coordinates $\times$ spatial transformations which are used for the assembly of nucleotides.

The use of large domains allows to cover a larger fraction of the conformational space and improves the precision of generated models. However, the size of the explored conformational space grows exponentially with the size of the domains. Assuming that the complexity of checking any given set of input constraints requires a constant amount of computer resources, then the complexity of the backtracking algorithm can be approximated with the use of three parameters: the number of variables, N; the size of the domains, M; and a measure of constraint efficiency, p, that could be thought of as a probability that any extension by one value of a partial solution will be consistent with the input constraints, and this at any time during the simulation. A probabilistic model based on these three parameters has been proposed by Haralick (Haralick 1980). In Haralick's model, M, N and p are constant values so that the expected number of partial structures during the simulation is given by $\sum_{i=1}^{N} M^i p^{(i-1)(i-2)/2}$; or M^N complete structures when $p = 1$.

In practice however, only N, the number of nucleotides is constant and known *a priori* of a simulation; M and p vary according to available constraints. In particular, it is extremely difficult to approximate p prior to run time. In practice, when p is small, combinatorial explosions are prevented and RNA structures are produced using reasonable amounts of computer resources. The theoretical size of the search tree for reproducing the yeast tRNA motif according to input constraints was 10^{28} (Major, Gautheret, & Cedergren 1993). Nevertheless, available constraints were used to sufficiently prune the search tree so that only half a million partial constructions were evaluated. This computation on currently available workstations takes only about five to 20 minutes. MC-SYM has been applied successfully to reproduce at the atomic level precision consensus structures such as the yeast $tRNA^{Phe}$, the UUCG tetranucleotide loop (Gautheret, Major, & Cedergren 1993) and the TYMV pseudoknot motif (Major *et al.* 1991), to provide support for the Sundaralingam model of the relative position and orientation of transfer RNAs in the ribosomal A and P sites (Easterwood *et al.* 1994), and using covariation data, to model the bound conformation of the Rev-binding element of HIV-1 (Leclerc, Cedergren, & Ellington 1994).

Tandem G.A mismatches

Non canonical base pairs, and in particular those involving a guanine and an adenine (G.A), play important structural and functional roles in both deoxyribonucleic acid (DNA) and RNA. There are four different G.A base pair patterns involving two hydrogen bonds (H-bonds). Following Saenger's notation, the four patterns are: VIII (A:N6-H...O6:G; A:N1...H1:G), IX (A:N6-H...O6:G; A:N7...H1:G), X (A:N6-H...N3:G; A:N1...H-N2:G) and XI (A:N6-H...N3:G; A:N7...H-N2:G) (Saenger 1984).

Several structures containing tandem G.A mismatches have been studied by nmr spectroscopy, X-ray crystallography and molecular modeling (see Table 1). A comparative sequence analysis reveals that tandem G.A mismatches occur frequently in the internal loops of 16S and 23S rRNA (Gautheret, Koonings, & Gutell 1994). The most frequent motifs are $\begin{smallmatrix}5'-G-A-3'\\3'-A-G-5'\end{smallmatrix}$, $\begin{smallmatrix}5'-A-A-3'\\3'-A-G-5'\end{smallmatrix}$ and $\begin{smallmatrix}5'-A-A-3'\\3'-G-G-5'\end{smallmatrix}$. On the basis of covari-

Duplex	Structural features	Reference
r(GGCGAGCC)$_2$	cross-strand stacking AG(XI)	SantaLucia & Turner (1993) *Biochemistry* **32**:12612.
d(ATGAGCGAATA)$_2$	cross-strand stacking AG(XI)	Li, Zon & Wilson (1991) *Proc. Natl. Acad. Sci. (USA)* **88**:26.
d(GCGAATAAGCG)$_2$	cross-strand stacking AG(XI); AA(41)	Maskos et al. (1993) *Biochemistry* **32**:3583.
Hammerhead ribozyme	cross-strand stacking	
d(-AG-)$_2$	AG(XI)	Katahira et al. (1993) *Nucleic Acids Res.* **21**:5418.
r(-AG-)$_2$†	AG(XI)	Pley, Flaherty & McKay (1994) *Nature* **372**:68.
d(CCAAGATTGG)$_2$†	AG(VIII)	Privé et al. (1987) *Science* **238**:498.
d([GA]$_n$)$_2$‡	AG(VIII); AG(IX)	Huertas et al. (1993) *EMBO J.* **12**:4029.

Table 1: A list of RNA and DNA duplexes for which nmr, X-ray crystallography† and molecular modeling‡ studies have been done to determine 3-D structure.

ation analysis, the $\begin{smallmatrix}5'-G-A-3'\\3'-A-G-5'\end{smallmatrix}$ motif is interchangeable with three motifs: $\begin{smallmatrix}5'-A-A-3'\\3'-A-G-5'\end{smallmatrix}$, $\begin{smallmatrix}5'-G-A-3'\\3'-A-A-5'\end{smallmatrix}$ and $\begin{smallmatrix}5'-A-A-3'\\3'-A-A-5'\end{smallmatrix}$, suggesting isosteric conformations.

Implementation

All simulations were performed using "ADJACENCY 1.0 3.5", that is, the distance between the terminal nucleotide O3' and the P atom of the next nucleotide was required to be no shorter than 1.0Å and no longer than 3.5Å. The following "GLOBAL" constraints, which describe the minimum distance in Å between pairs of atoms were used: P P 3.0; C1' C1' 3.0; PSE PSE 3.0; P C1' 2.5; P PSE 2.5; C1' PSE 2.5; C4 C4 2.0; N1 N1 2.0; N1 C1' 2.0; N1 C4 2.0; O2' O4' 2.0; O3' C5' 2.0; P C3' 2.0; O2' C3' 2.0; O3' C3' 2.0; C2 N7 1.5; and, C2 C2' 1.5. These constraints are used as a heuristic control for steric avoidance and, in absence of more detailed energy function, to allow for the computation of larger numbers of conformations in given amounts of time.

For each interchangeable motif, $\begin{smallmatrix}5'-G-A-3'\\3'-A-G-5'\end{smallmatrix}$, $\begin{smallmatrix}5'-A-A-3'\\3'-A-G-5'\end{smallmatrix}$, $\begin{smallmatrix}5'-G-A-3'\\3'-A-A-5'\end{smallmatrix}$ and $\begin{smallmatrix}5'-A-A-3'\\3'-A-A-5'\end{smallmatrix}$, an MC-SYM script was defined and used to generate libraries of conformations. The non canonical purine-purine base pairs from (Saenger 1984) were implemented and systematically tested for G.A and A.A. To find isosteric substitution for G.A, the A.A base pairs involving one H-bond were implemented and tested as well; using the following Arabic numbering notation: AA[35,36,37,38,39,40,41] (see Figure 2).

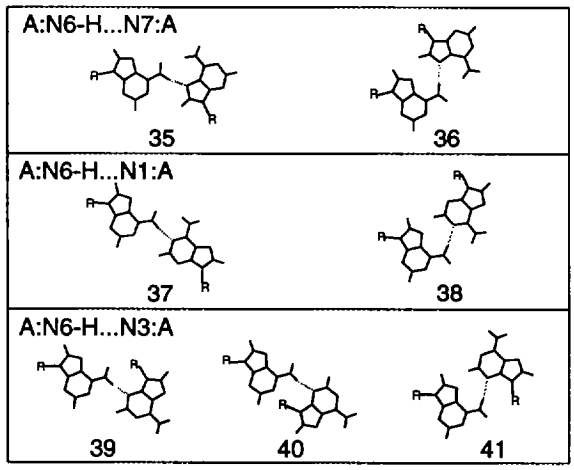

Figure 2: A.A base pairing patterns composed of one H-bond as implemented in MC-SYM.

The general script format is as follow:

```
; 5'-X-A-3'
; 3'-A-Y-5'
SEQUENCE
   rX   1   reference       <type>
   rA   2   <stack>     1   <type>
   rY   3   <pair>      2   <type>
   rA   4   <pair>      1   <type>
```

The *rX* identifies the type of nucleotides. The integers on the right side of the nucleotide types are references in the sequence. The next column indicates the possible spatial transformations. The integers on the right side of the spatial transformations are references to the nucleotides used in the application of the specified transformations. Finally, the last column indicates the type of nucleotide conformations that will be assigned during the simulations. The script can be read as follow: Use nucleotide 1 as the reference. Stack the base of A2 on the base of nucleotide 1. Append nucleotide 3 by trying all possible base pairings to form the pair A2.Y3. Append A4 by trying all possible base pairings to form the pair A4.X1.

In this script, *rX* and *rY* were substituted according to the value of *X* and *Y* in the sequence, that is, *rA* or *rG*. The *reference* spatial transformation is a special one indicating a referential for the construction. The *stack* transformations ensured for the stacking of bases 1 and 2 and included cross-strand stacking. The *pair* transformations apply the non canonical base pairing geometries described above. The geometries vary with the type of nucleotides *X* and *Y*: AG_pair or AA_pair. The conformation of all nucleotides were assigned from a sampling of sugar puckers (C2'- and C3'-endo). In addition, all guanines were assigned a sample of conformations with *anti* torsion angle around the glycosyl bond. No such constraint were applied to the adenines so that *syn* torsion angles around the glycosyl bond were also assigned.

Results

For the tandem A.G, MC-SYM generates 893 structures that were classified into seven classes combining the following A.G base pairing patterns: IX-IX, VIII-IX, IX-VIII, VIII-VIII, IX-XI, X-XI and XI-XI (see Table 2). Sugar pucker modes and torsions around the glycosyl bond were then used to regroup the structures into 78 subclasses. The class VIII-VIII represents near 50 percent, and the class XI-XI near 30 percent of all generated structures. The adenines are in *syn* and the guanines in *anti* conformations in all structures of class IX-IX. In the symetrical VIII-IX and IX-VIII tandems, the adenines involved in type IX base pairs are in *syn* conformations. All other bases are in *anti* conformations. In the tandems of class VIII-VIII, IX-XI and XI-XI, all bases are in *anti* conformations. Finally, in the X-XI tandem the adenine in the type X base pair is in *syn* conformation. The other bases are in *anti* conformations. The interstrand stacking observed in the nmr structures of the class XI-XI tandem has been reproduced properly (see Figure 3a).

Molecule	Class	Subclass	Occurence
	1•4-2•3	(#)	(%)
5'-G-A-3' (1,2) / 3'-A-G-5' (4,3)	IX-IX	1	0.3
	VIII-IX	4	3.4
	IX-VIII	7	11.2
	VIII-VIII	34	49.8
(893)	IX-XI	8	5.0
	X-XI	4	1.7
	XI-XI	20	28.6
5'-G-G-C-G-A-G-C-C-3' (1,2) / 3'-C-C-G-A-G-C-G-G-5' (4,3)	VIII-IX	6	8.3
	IX-VIII	18	21.4
	VIII-VIII	60	69.7
	IX-XI	2	0.4
(9,072)	XI-XI	1	0.1

Table 2: Summary of MC-SYM simulations. The numbers in parentheses indicate the number of models generated. Occurence percentages are relative to the number of models generated in each case.

The library of tandems has been used to model the r(GGCGAGCC)$_2$ duplex of SantaLucia et al. (SantaLucia & Turner 1993). MC-SYM generates 9,072 such models. The models were grouped into five structural classes: VIII-IX, IX-VIII, VIII-VIII, IX-XI and XI-XI, and further divided in 87 subclasses. Near seventy percent of the models fall into the class VIII-VIII. The class XI-XI, observed by nmr spectroscopy, represents only about 0.1 percent of the models. This class contains only one subclass; with a maximum rms deviation between all-atoms (except hydrogens) of 0.01Å. The rms deviation between the MC-SYM models and one of the nmr structures is 3.9Å. The rms deviation between a refined MC-SYM model (by applying molecular mechanics energy minimization) and the nmr model is 1.5Å. The most deviation comes from the backbone atoms in the tandem which are not symmetric in MC-SYM models. Another important characteristic of the solution model, which is no well reproduced in the MC-SYM models or their refinements, is the formation of H-bonds between the adenine-amino protons and the O2' of the guanine, and between the guanine-amino protons and the adenine-phosphate oxygens.

In the context of comparative sequence analysis, two series of isosteric conformations for the variations of $\begin{smallmatrix}5'-G-A-3'\\3'-A-G-5'\end{smallmatrix}$, that is, $\begin{smallmatrix}5'-A-A-3'\\3'-A-G-5'\end{smallmatrix}$, $\begin{smallmatrix}5'-G-A-3'\\3'-A-A-5'\end{smallmatrix}$ and $\begin{smallmatrix}5'-A-A-3'\\3'-A-A-5'\end{smallmatrix}$, have been identified. The A.G-XI;A.A-41 (Figure 3b) and A.A-41;A.A-41 (Figure 3c) tandems are isosteric to the A.G-XI;A.G-XI (Figure 3a) tandem observed by nmr and suggested in (Gautheret, Koonings, & Gutell 1994). Interestingly, other isosteric con-

formations were identified by our method. The A.G-VIII;A.G-VIII (Figure 3d), A.G-VIII;A.A-37 (Figure 3e) and A.A-37;A.A-37 (Figure 3f) are also isosteric.

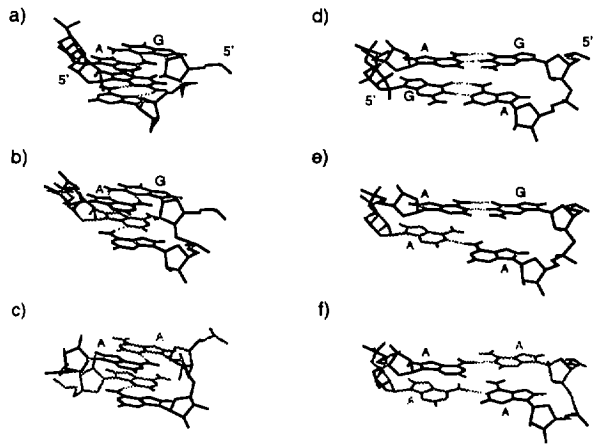

Figure 3: Isosteric motifs. Three-dimensional structures generated by MC-SYM for the tandem of class a) XI-XI, b) XI-41, c) 41-41, d) VIII-VIII, e) VIII-37, and f) 37-37. Adenines substituting guanines are in gray.

Conclusion

The goals of this research were to demonstrate the accuracy of MC-SYM in reproducing experimental results for small RNA motifs and to show how the modeling tool can be used to generate libraries of conformations for small RNA motifs. Although it would be very difficult to search the conformational space of RNA exhaustively, the results presented here indicate that MC-SYM is accurate in generating libraries of structural motifs from sequence information. The finding of isosteric conformations to the A.G-VIII;A.G-VIII tandem suggests that this conformation is also possible in the context of 16S and 23S rRNAs. Energy evaluation and fitness of precise chemical data would be necessary to distinguish among the two possibilities.

Acknowledgments

We thank Dr. D. Gautheret for providing A.A base pairing geometries involving a single H-bond. We thank Dr. J. Santa Lucia Jr. for providing atomic 3-D coordinates of the nmr structure of the r(GGCGAGCC)$_2$ duplex. FM is a fellow of the Canadian Genome Analysis and Technology program and the Medical Research Council (MRC) of Canada.

References

Altman, R. B.; Weiser, B.; and Noller, H. F. 1994. Constraint satisfaction techniques for modeling large complexes: application to the central domain of 16S ribosomal RNA. In Altman, R.; Brutlag, D.; Karp, P.; Lathrop, R.; and Searls, D., eds., *Proceedings, second international conference on intelligent systems for molecular biology.* Menlo Park, CA 94025: AAAI Press. 10–18.

Altman, R. B. 1993. Probabilistic structure calculations: A three-dimensional tRNA structure from sequence correlation data. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings, first international conference on intelligent systems for molecular biology.* Menlo Park, CA 94025: AAAI Press. 12–20.

Battiste, J. L.; Tan, R.; Frankel, A. D.; and Williamson, J. R. 1994. Binding of an HIV Rev peptide to Rev responsive element RNA induces formation of purine-purine base pairs. *Biochemistry* 33:2741–2747.

Cheong, C.; Varani, G.; and Tinoco, I. J. 1990. Solution structure of an unusually stable rna hairpin, 5'ggac(uucg)gucc. *Nature* 346:680–682.

Easterwood, T.; Major, F.; Malhotra, A.; and Harvey, S. 1994. Orientation of transfer RNA in the ribosomal A and P sites. *Nucleic Acids Res.* 22:3779–3786.

Gautheret, D.; Koonings, D.; and Gutell, R. 1994. A major family of motifs involving G.A mismatches in ribosomal RNA. *J. Mol. Biol.* 242:1–8.

Gautheret, D.; Major, F.; and Cedergren, R. 1993. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol.* 229:1049–1064.

Haralick, R. M. 1980. Increasing Tree Search Efficiency for Constraint Satisfaction Problems. *Artificial Intelligence* 14:263–313.

Heus, H., and Pardi, A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* 253:191–194.

Hilbert, M.; Bohm, G.; and Jaenicke, R. 1993. Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* 17:138–151.

Leclerc, F.; Cedergren, R.; and Ellington, A. D. 1994. A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nature: Structural Biology* 1:293–300.

Major, F.; Turcotte, M.; Gautheret, D.; Lapalme, G.; Fillion, E.; and Cedergren, R. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* 253:1255–1260.

Major, F.; Gautheret, D.; and Cedergren, R. 1993. Reproducing the three-dimensional structure of a transfer RNA molecule from structural constraints. *Proc. Natl. Acad. Sci. (USA)* 90:9408–9412.

Malhotra, A.; Tan, R. K.; and Harvey, S. C. 1990. Prediction of the three-dimensional structure of es-

cherichia coli 30s ribosomal subunit: A molecular mechanics approach. *Proc. Natl. Acad. Sci. (USA)* 87:1950–1954.

Michel, F., and Westhof, E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216:585–610.

Pan, T., and Uhlenbeck, O. C. 1992. A small metalloribozyme with a two-step mechanism. *Nature* 358:560–563.

Pley, H. W.; Flaherty, K. M.; and McKay, D. B. 1994. Three-dimensional structure of a hammerhead ribozyme. *Nature* 372:68–74.

Privé, G. G.; Heinemann, U.; Chandrasegaran, S.; Kan, L.-S.; Kopka, M. L.; and Dickerson, R. E. 1987. Helix geometry, hydratation and G.A mismatch in a b-dna decamer. *Science* 238:498–504.

Saenger, W. 1984. *Principles of Nucleic Acid Structure*. New-York: Springer-Verlag.

SantaLucia, J. J., and Turner, D. 1993. Structure of (rGGCGAGCC)$_2$ in solution from NMR and restrained molecular dynamics. *Biochemistry* 32:12612–12623.

Wimberly, B.; Varani, G.; and Tinoco, I. J. 1993. The conformation of loop E of eukaryotic ribosomal RNA. *Biochemistry* 32:1078–1087.