Pacific Symposium on Biocomputing 15:216-227(2010)

PREDICTING RNA STRUCTURE BY MULTIPLE TEMPLATE HOMOLOGY MODELING

SAMUEL C. FLORES[†*], YAQI WAN[°*], RICK RUSSELL[°], RUSS B. ALTMAN[†]

[†]*Bioengineering Department, Stanford University, Clark Center S231, 318 Campus Drive, Stanford, California 94305-5444, USA*

[°]*Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology University of Texas at Austin, 1 University Station A4800, 2500 Speedway Austin, Texas 78712, USA*

Despite the importance of 3D structure to understand the myriad functions of RNAs in cells, most RNA molecules remain out of reach of crystallographic and NMR methods. However, certain structural information such as base pairing and some tertiary contacts can be determined readily for many RNAs by bioinformatics or relatively low cost experiments. Further, because RNA structure is highly modular, it is possible to deduce local 3D structure from the solved structures of evolutionarily related RNAs or even unrelated RNAs that share the same module. RNABuilder is a software package that generates model RNA structures by treating the kinematics and forces at separate, multiple levels of resolution. Kinematically, bonds in bases, certain stretches of residues, and some entire molecules are rigid while other bonds remain flexible. Forces act on the rigid bases and selected individual atoms. Here we use RNABuilder to predict the structure of the 200-nucleotide *Azoarcus* group I intron by homology modeling against fragments of the distantly-related *Twort* and *Tetrahymena* group I introns and by incorporating base pairing forces where necessary. In the absence of any information from the solved *Azoarcus* intron crystal structure, the model accurately depicts the global topology, secondary and tertiary connections, and gives an overall RMSD value of 4.6 Å relative to the crystal structure. The accuracy of the model is even higher in the intron core (RMSD = 3.5 Å), whereas deviations are modestly larger for peripheral regions that differ more substantially between the different introns. These results lay the groundwork for using this approach for larger and more diverse group I introns, as well for still larger RNAs and RNA-protein complexes such as group II introns and the ribosomal subunits.

1. Introduction

RNA plays pervasive roles in gene regulation and expression. Messenger RNA provides the template for protein synthesis but also forms structures to regulate that synthesis (1). MicroRNAs inhibit protein production by promoting degradation of their targeted mRNA transcripts or stalling of their translation (2, 3). Even the remarkable machinery that synthesizes proteins, the ribosome, is composed primarily of RNA. Further, recent genomics approaches have indicated that much of the human genome is transcribed and most of it is not translated into protein, suggesting that many more functions of RNA are yet to be discovered (4, 5). The diversity of functional roles for RNA has profound implications for the early development of life, and the elucidation of these functions holds the promise of novel treatments for human diseases (6).

However, our understanding of RNA structure and function is continually hampered by a persistent lack of structural coordinates. Part of the challenge arises from the experimental difficulty of crystallizing a highly charged, very flexible molecule with a dearth of the distinctive surface features needed for specific crystal packing (7). Compounding the problem, many structured RNAs can adopt alternative conformations at equilibrium or as long-lived kinetic traps during folding (8, 9). Theoretical approaches are also challenged by these features, both by the delicate energetic balance between alternative conformations and by the long times required to equilibrate during folding.

The structures of relatively small RNAs can often be predicted by one of several methods. Fragment Assembly of RNA (FARNA) assembles structures by sampling trinucleotide fragments from a database and screens these structures using a coarse grained potential that favors base pairing and stacking geometries (10). Similarly, MC-Sym samples four-nucleotide cycles from a database that are consistent with known base pairing contacts and

[*] These authors contributed equally to the work.

progressively builds up structure (11). Both of these methods, however, have computer time requirements that scale poorly with size, and rely on fragment databases that include only limited diversity. They therefore have not been shown to predict the structure even of molecules as large as tRNA (at ~75 nucleotides), except when using a fragment library that contains tRNA (12). Discrete Molecular Mechanics (DMD) has a simplified potential for RNAs, which are represented by three pseudoatoms per nucleotide; it can solve the structure of tRNA but larger molecules remain out of reach. The Nucleic Acid Simulation Tool (NAST) uses one pseudoatom per nucleotide to represent RNA structure and can fold the (~150 nt) P4/P6 domain of the *Tetrahymena* group I intron (13). However, its force field is too coarse to discriminate the native state from decoys at larger size scales.

Homology modeling can be used to predict the structure of larger molecules when structural homologs are available. The first step is to obtain a correspondence between residues in the molecule to be modeled and a template; if the sequence identity is low this process must be done manually. The next step is to geometrically align residues from the model onto corresponding residues in the template. Lastly, the structure of any inserted regions must be solved and deletions must be closed. Kevin Sanbonmatsu and collaborators have threaded the *E. coli* 16S ribosomal RNA onto a template from *Thermus Thermophilus*. The model and template have 75% sequence identity, considerably higher than that between *Azoarcus* and *Twort* ribozymes (<50%). The only insertions are in nine hypervariable regions that are not in contact with each other. These were dealt with by adding fragments from additional structures. Since the model does not have the flexibility needed to align and form long range contacts, it would be difficult to build structures that are interconnected in regions with no single structural homolog, as occurs with the *Azoarcus* ribozyme. Further, there is no code distributed for this method, leaving the technique to be applied by computational experts (14).

There is therefore a need for a homology modeling program that can *1)* structurally align corresponding residues while allowing close user control over the threading, *2)* incorporate templates from a third molecule or molecules, *3)* solve the structure of connecting regions with no template by enforcing sterics, chemistry, and base pairing while allowing flexibility and *4)* be accessible to the experimentalist. In the current work we describe a multi-resolution modeling approach (see Background) that does all of this.

Requirement *1* is met by applying forces which align the threaded and template bases. These are specified by the user just like the base pairing forces described below. Requirement *2* is met by adding more templates and connecting them to the threaded molecule. The molecules are completely rigid, a feature of our multi-resolution modeling (MRM) approach.

Requirement *3* is the ability to predict structures of connecting regions by applying steric exclusion and base-pairing forces while maintaining bond lengths and angles. Our method requires secondary and tertiary base pairing contacts, which are provided by the user in a simple format. Each of these contacts becomes a term in our force field, with force and torque components, which act to bring the paired bases into the indicated geometry. Simulated annealing is then used to escape kinetic traps or local energy minima. The kinematics treats each backbone atom as an independent body and the bases as rigid units. In parallel, the base pairing forces act on a single atom per base, and the sterics are treated with contact spheres on a few atoms per nucleotide. The parallel treatment of forces and mobilities at different levels of resolution is another MRM aspect of our method. RNABuilder can solve the structures of small connecting regions without a template and does not compute interactions between all near neighbors, thus addressing the scaling issues inherent in both fragment assembly and Molecular Dynamics(MD)-like methods.

Requirement *4* is met by providing an easy-to-use interface that was used in this work to specify all parameters of our model, including the sequences, base-pairing contacts, and correspondence between the model and templates. A single input file is prepared by the user to provide the RNA sequences, base-pairing contacts, and simulation

parameters. A tutorial is provided (*Distribution* section), which includes example input files to use as a starting point for the user's own runs. Although no significant computational skill is required, modeling decisions require molecule-specific knowledge which an experimentalist would be expected to have about his or her system.

Here, we use RNABuilder to model the structure of a large, multi-domain RNA, the 200-nucleotide *Azoarcus* group I intron, by homology modeling with multiple templates. Group I introns have been powerful model systems for understanding RNA structure, folding, and function since they were discovered as the first catalytic RNAs more than 25 years ago (9, 15-18). By comparing our model to the solved crystal structure of the intron, we show that the model correctly captures the domain structures, connections and topology and gives an overall RMSD of 4.6 Å.

2. Background

2.1. *Classification of multi-resolution modeling (MRM) techniques*

Multi-resolution modeling (MRM) has emerged as a useful paradigm. MRM refers to the modeling of molecules at coarse grained resolution, with direct or indirect connection to the corresponding atomistic model (in which each atom is an independent body). According to the Ayton, Noid, and Voth classification (19), such schemes can be either serial or parallel. In a serial scheme, the parameters and form of the coarse grained model can be developed from the atomistic one (S-A), from various sources of knowledge (S-B), or from thermodynamic data (S-C). In a parallel scheme, the coarse and atomistic models can interact (P-1) or run concurrently under a resolution exchange methodology (P-2). RNABuilder is an MRM method of type P-1, since the kinematics (which in certain places are atomistic, and in other places very coarse) and the forces (mostly coarse grained) form part of a single system.

2.2. *Multibody dynamics*

Internal coordinate multibody dynamics is a method for simulating molecules. Instead of storing the Cartesian coordinates of each atom as would be done in a conventional MD simulation, one represents the state of the molecule via its natural "internal coordinates", such as bond lengths and torsional angles (20). It then becomes trivial to constrain any internal coordinate: one simply keeps it fixed, only integrating the unconstrained degrees of freedom. Internal coordinate multibody dynamics is much more difficult to implement than traditional MD, since it is necessary to transform atom locations to Cartesian coordinates before calculating forces, to transform the forces back to internal coordinates again before integrating, and to calculate the effective inertia of each coordinate based on the current state of the molecule at each time step; however there exist efficient linear time algorithms for performing these operations. When a system is highly constrained, internal coordinate multibody dynamics is far more efficient than Cartesian MD. In fact, the more highly constrained the system is, the more efficient it becomes, in direct contrast to Cartesian constraint algorithms which become less efficient as constraints are added (21).

2.3. *Simbody and Molmodel*

Our RNABuilder package is written using the Simbody (22) internal coordinate mechanics library and its molecular mechanics extension, Molmodel, both available from SimTK.org. Simbody includes variable step size integrators (23), which continually attempt to maximize the integration step size without exceeding error tolerances, leading to significant time savings as well as stable behavior when dealing with large forces. Molmodel permits us to rigidify all template molecules and part of the modeled structure saving the computational expense often associated with modeling large structures. RNA bases can be modeled as rigid bodies and so base pairing forces and torques can be applied to the base rather than to individual atoms, reducing the number of calculations to be performed. Lastly, Simbody provides collision-detecting Contact spheres (24), which we use on selected atoms as an approximate treatment of sterics.

2.4. *RNABuilder force field, mobilizers, and constraints*

The coarse grained force field used in this work consists of forces and torques which act to bring the interacting bases into the base pairing geometry specified by the user using the *baseInteraction* commands. No forces act between bases unless specified by the user, except stacking forces which are automatically added to helices. The first base has an attachment frame which is part of the glycosidic nitrogen body but which is often located several angstroms from the nucleus, and which has a specified relative angular orientation. The second base has only a body frame located at the glycosidic nitrogen nuclear center and rotated such that the x-axis lies along the glycosidic bond axis and the z-axis lies normal to the base plane and points in the 3' direction, assuming helical geometry. The forces act to pull the attachment and body frames together in cartesian space, and the torques act to align them rotationally. Since the bases are rigid by default this is sufficient to bring all atoms into a desired base pairing geometry. Parameterization of the force field thus consists primarily of determining the translation and rotation of the attachment frame that will result in the base pairing geometry; a program is provided to compute these parameters but most users will simply use the distributed parameters. These parameters include those corresponding to all base pairs catalogued by Leontis et al (including the WatsonCrick) (25), plus stacking and a Superimpose interaction used in threading to align the threaded with the template base. See Methods for enforcement of these interactions.

The geometric relationship between attachment and body frames is enforced by means of a potential that has a strength parameter cutoffPotential (user adjustable in the RNABuilder parameter file) based loosely on base pairing enthalpies (26, 27). It also has a range parameter cutoffRadius (set by the user in the RNABuilder input file) based on ranges reported for base pairing and stacking forces (10, 27). RNABuilder detects runs of three or more consecutive WatsonCrick base pairs and automatically enforces helical geometry. The potential is harmonic at close range and inverse at long range. The requirement that the potential and its derivative match at cutoffRadius leads to the following form:

$$U_{translational}(r) = \begin{cases} \frac{cutoffPotential}{2} \cdot (3 - \frac{r^2}{cutoffRadius^2}), 0 \le r < cutoffRadius \\ \frac{cutoffPotential \cdot cutoffRadius}{r}, r \ge cutoffRadius \end{cases}$$

The force is obtained by differentiation. These quantities are plotted in Figure 1 below.

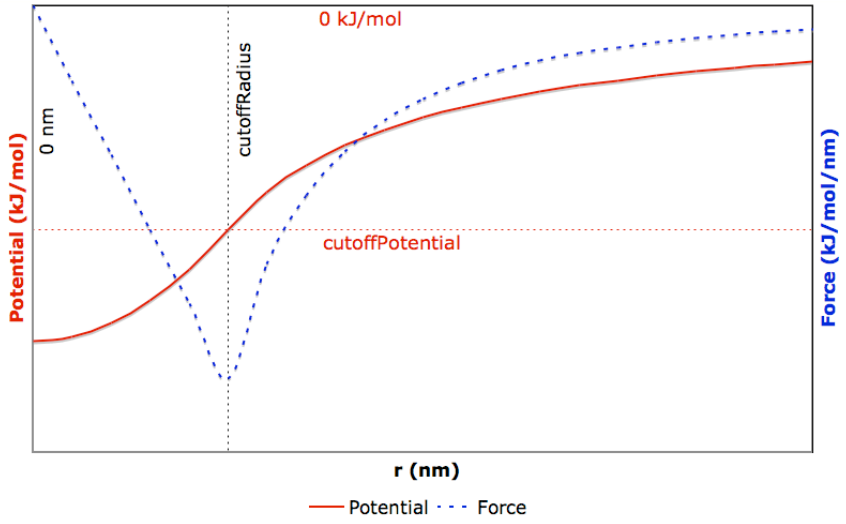

Figure 1. Potential and force as a function of distance from *attachment frame* of residue 1 to *body frame* of residue 2.

We provide a *contact* force, which can attach spheres to selected atoms to prevent steric clashes. These spheres interact repulsively when they overlap. The HardSphere keyword specifies that spheres be added to 2-3 atoms per residue, for any specified range of residues. We used this on selected residues to push apart overlapping strands.

RNABuilder also provides a series of *mobilizer* commands which set the flexibility of the molecule. The default mobilizer rigidifies all base bonds; most backbone bond lengths and angles are fixed but dihedral angles are allowed to vary. This guarantees that all bonds will have reasonable local geometry. In this work we use the 'Rigid' mobilizer to rigidify the template molecules, and parts of the threaded molecule which at certain stages were considered to be converged. The Rigid mobilizer reduces computational expense since it simply removes certain degrees of freedom rather than imposing constraint equations which must then be solved.

Lastly, RNABuilder includes a *constraint* called 'Weld' which fixes the C3' carbon of any residue to that of any other. We used this where two separate chains had been rigidified and needed to be held together. A constraint comes at a cost since it adds an equation that must be solved in Simbody.

3. Methods

The *Azoarcus* group I intron sequence was aligned and threaded to the *Twort* intron (Fig. 2) (28). The RNAs include secondary structure features that are numbered by convention and designated P (helix), J (junction), or L (loop). Direct correspondences are present over most of the length of the RNA, allowing straightforward threading of secondary elements P3, P4, P6, and P7, as well as several internal lops and single-stranded joining segments (J3/4, J6/7, J8/7, J4/5, and J6/6a). However, in regions that are sufficiently different between the two RNAs that direct correspondences are not present, alternative methods were used. Two 'tetraloop-receptor' tertiary interactions were threaded against a fragment of the Tetrahymena intron that includes the same type of interaction. For three short connecting segments, no existing group I intron structures were identified with corresponding sequences in the same positions and with the same numbers of nucleotides, and these segments were therefore left unthreaded (shown in black in Fig. 2). Similarly, two hairpin loops were not threaded because structures with the same sequences were not available.

3.1. *Modeling the P1-P2 stacked helix*

The P1 helix, which includes the strand that is cleaved during splicing, is three base pairs in the *Azoarcus* intron but four base pairs in the *Twort* intron. However, this length difference is compensated by a difference in P2, which is one base pair longer in the *Azoarcus* intron. Therefore, the first base pair of the *Azoarcus* P2 was threaded to the fourth base pair of P1 in the *Twort* structure. Because the final base pair of the *Azoarcus* P2 element was modeled onto P1, but the joining sequence J2/3 emanates from P2, J2/3 was not threaded.

3.2. *Modeling the P9 region*

From its sequence, the *Azoarcus* intron has the potential to form a P9.0 helix consisting of two G-C base pairs, but on the 5′-side there are three consecutive Gs, and it is not clear from the sequence which two form these base pairs (see Fig. 2). We modeled base pairs by the two 5′-most Gs, which allowed threading of the single nucleotide linking P7 and P9.0. From P9.0 to P9 on the 5′-side, there are two unpaired nucleotides in the *Azoarcus* intron sequence. We threaded these nucleotides into P9, where they could form one canonical and one non-canonical base pair, generating a P9 helix that would be the same length as in the *Twort* intron. The *Azoarcus* intron has three unpaired nucleotides connecting P9 to P9.0 (J9/9.0), whereas the *Twort* intron has seven unpaired nucleotides connecting in the corresponding junction. These three nucleotides were threaded onto the three nucleotides immediately upstream of P9.0 in the *Twort* intron.

3.3. *Modeling tetraloop-receptor interactions using additional templates*

It was not possible to thread two tetraloop-receptor interactions of the *Azoarcus* intron. One of these sites was disordered in the *Twort* intron crystal structure and the other site has a different type of receptor (29, 30). We threaded both of these interactions using the structure of the P4-P6 domain of the *Tetrahymena* ribozyme (31), which has the same type of tetraloop and receptor as the *Azoarcus* intron. After completing an initial round of all of the modeling steps (see below), we observed that one of the loops, L2, had moved so that it approached its receptor too closely to be physically reasonable. We therefore reinforced the correct conformations in this contact by again threading these regions to the *Tetrahymena* intron template (see Results).

3.4. *Hairpin loops*

Some hairpin loops could not be threaded to the *Twort* intron because of natural or engineered differences. Two such hairpins, P5a and P8a, were threaded to sequences in the *Tetrahymena* intron as described above. A third, P6a, was omitted and modeled as a blunt-ended helix.

3.5. *Using RNABuilder to perform threading*

We began by threading the flexible *Azoarcus* group I ribozyme onto the *Twort* ribozyme using the RNABuilder commands described in the Background section. There is a gap in the *Twort* structure where the distal end of P5 is deleted; accordingly we modeled this template using two strands which were then rigidified and welded to each other. We used the 'Superimpose' twoTransformForce to pull together bases in *Azoarcus* and *Twort* that were aligned as described above. Of the residues that have no equivalent in *Twort*, some were left to be folded by threading to additional templates as described above. Others were in known helical regions and modeled by applying 'WatsonCrick' twoTransformForces. The 'HardSphere' contact was applied only to residues where a steric clash occurred.

The model we obtained in this step was a global 3D structure with P1/2, P4-P6, P3-P7 and P9 formed exactly as *Twort* ribozyme but with several other portions not yet folded. The latter included the tetraloop-receptor in L2-P8, especially the receptor part in P8, since there is no template in *Twort* for this type of receptor. *Azoarcus* L9-P5 is similar to *Twort* but as mentioned the top of P5 stem including half of the receptor was distorted and deleted from the crystal structure. Lastly, the distal end of P6a in the *Azoarcus* intron construct used for crystallization has an engineered U1A loop, which of course is not present in the *Twort* intron or in the natural *Azoarcus* intron. We introduced a chain break in the model between nucleotides 109 and 110 to circumvent the engineered loop.

To model L2-P8 and L9-P5, in the second stage, we introduced as additional templates two copies of the L5b and P6 fragments from the *Tetrahymena* ribozyme. In each copy, L5b and P6 were rigidified using the 'Rigid' mobilizer and welded together using the 'Weld' constraint. Then 'Superimpose' baseInteractions were used to attach one copy each to the ends of the *Twort* P8 and P5 helices. With these prostheses the template now had the desired motifs in L2-P8 and L9-P5. The appropriate *Azoarcus* intron nucleotides were then threaded onto the corresponding nucleotides of the *Tetrahymena* intron fragment.

4. Results

To use homology modeling to predict the structure of the *Azoarcus* group I intron, we first aligned the secondary structure of the *Azoarcus* intron, established by comparative sequence analysis, to a version of the *Twort* secondary structure that was validated by its crystal structure (28) (Fig. 2). Although the crystal structure of the *Azoarcus* intron has been solved (32, 33), we did not use any information from this structure in the alignment or in the subsequent modeling. Despite representing different subgroups – the *Azoarcus* is designated as a IC3 intron and the *Twort* intron is a IA2 intron – the core elements of secondary structure and their connections are conserved (34). Further, many of these secondary structure elements and connections are identical in length between the two introns

and were aligned in a straightforward manner. Although the *Twort* intron includes some helical elements that are not present in the *Azoarcus* intron, the *Azoarcus* intron is not as large and complex as many group I introns, and all of its helical elements are present in the *Twort* intron.

Nevertheless, some structural features differ significantly between the two introns, such that it was not possible to directly thread certain local regions to the *Twort* intron structure (Fig. 3; see Methods). Most notably, two tertiary contacts in the *Azoarcus* intron are formed by canonical tetraloop-receptor interactions between a GAAA tetraloop and a receptor that includes an internal loop and has been termed the 11-nucleotide receptor (29). One of these contacts is between L2 and P8 and the other is between L9 and P5. Although the *Twort* intron has tetraloop-receptor interactions at equivalent positions, the interacting partners are of different structural classes. Instead of being the 11-nucleotide receptor, one of the receptors consists of two Watson-Crick base pairs, and the tetraloop for this receptor is GUAA instead of GAAA. The other receptor, in P5, includes an internal loop and interacts with a GAAA tetraloop. However, it is a non-canonical version of the 11-nucleotide receptor, and it could not be used as a template regardless because it was disordered in the crystal structure (28).

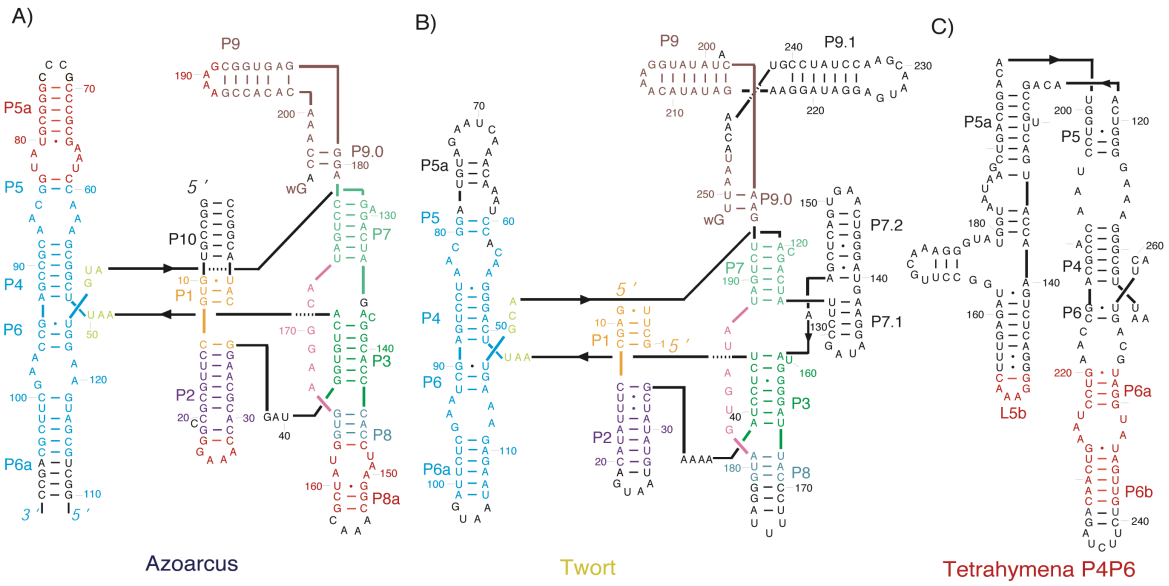

Figure 2. The secondary structures of A) the *Azoarcus* intron, B) the *Twort* intron, and C) the P4-P6 domain of the *Tetrahymena* intron. The domains are color-coded, with like colors indicating a correspondence between the *Azoarcus* model and the *Twort* or *Tetrahymena* intron templates. This correspondence was used as the basis for threading. Note that the tetraloop-receptor structure from the interaction of L5b with J6a/6b of the *Tetrahymena* intron (orange) was used as a template for both tetraloop-receptor contacts within the *Azoarcus* intron. Regions that were not threaded are shown in black.

To model these tetraloop-receptor interactions, we took advantage of the presence of the canonical 11-nucleotide receptor in the P4-P6 domain from the *Tetrahymena* intron, whose structure has been solved by x-ray crystallography (31). We aligned the *Tetrahymena* intron fragment to the *Twort* intron structure by using the base pairs adjacent to the receptor motif and then threaded the tetraloop and receptor sequences from the *Azoarcus* intron onto the corresponding nucleotides within the *Tetrahymena* intron fragment (Fig. 3; see also Fig. 2).

The final model of the *Azoarcus* intron was evaluated by superimposing it against the structure of the intron determined by crystallography (Fig. 4). The model has an RMSD with respect to the crystallographic coordinates of 4.59 Å (Table 1). The topology is correct throughout the intron, and the placement of helices is also largely correct. The closest overall agreement is found in the core region, with an RMSD value of 3.54 Å. The active site has a slightly higher RMSD value of 3.70 Å.

The peripheral regions also give a good overall agreement with the structure, as illustrated in Fig. 4. Nevertheless, larger differences were observed in the periphery than in the core. The receptor within P8, which was modeled using the *Tetrahymena* intron fragment, was positioned correctly and maintained the correct local structure, as expected from the known structural conservation of this tetraloop-receptor motif (31-33). However, a portion of P2 is shifted downward along its axis relative to its position in the crystal structure. Because the interaction of the tetraloop with the receptor in P8 was enforced by threading directly to the *Tetrahymena* intron fragment, the shift of P2 results in a minor, local distortion of P2 close to its distal end (Fig. 5A). Intriguingly, the shift of P2 can also be seen for the *Twort* intron structure relative to the *Azoarcus* intron structure (data not shown), suggesting that the subtle structural mismatch in this region, relative to the *Azoarcus* intron, is inherent to the *Twort* intron template.

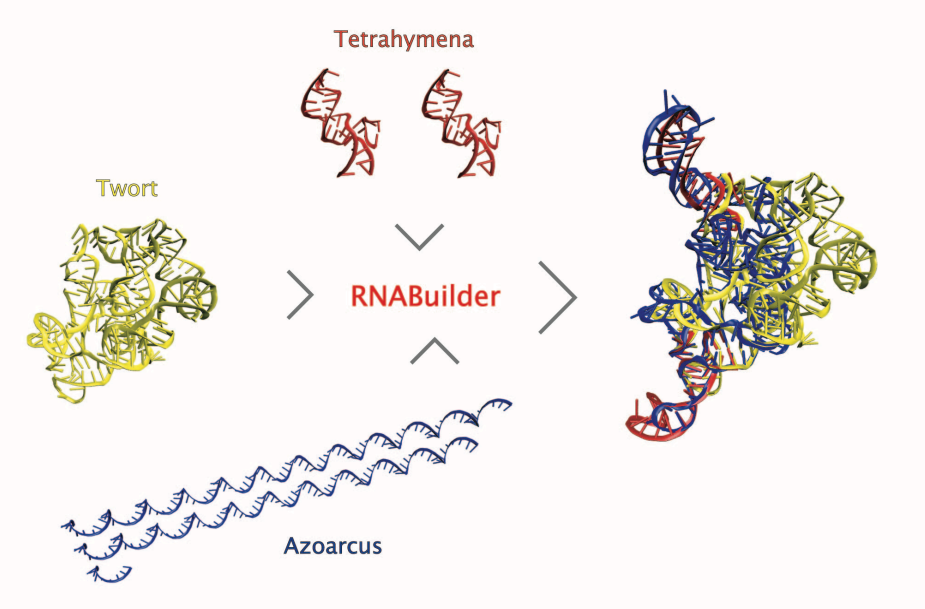

Figure 3. Modeling the *Azoarcus* intron (blue) by threading to fragments of the *Twort* and *Tetrahymena* intron structures. Three rigid fragments were used as templates: the nearly-complete intron from *Twort* (left, yellow) and the tetraloop receptor from the P4-P6 domain of the *Tetrahymena* intron (top, orange). The fragments corresponding to the *Azoarcus* intron were initially in extended conformations (bottom). The final model of the threaded *Azoarcus* intron, superimposed on the *Twort* and *Tetrahymena* template fragments, is shown at right.

Table 1

Region	Nucleotides	RMSD (Å)
Entire intron	4-206	4.59
Core: P1, P3, P4, P6, P7, P8, P8a	10-12, 41-59, 86-99, 120-153, 158-178, 206	3.54
Active site	8-12, 85-89, 126-132, 137, 168-181, 180-181, and 203-205, and substrate residues -3 to +2	3.70
Periphery: P2, P5, P5a, P6a, P9, P9.0	13-40, 60-84, 100-119, 154-157, 179-205	5.40

There are also small but significant differences in the P9 region. Notably, nucleotides G182 and A183, which were threaded into P9 of the *Twort* intron, do not form P9 pairs in the *Azoarcus* intron crystal structure (Fig. 5B). Instead,

they stack with the preceding nucleotides, forming an extension of P9.0 with non-canonical base pairs to A201 and A202. In large part because of this base-pairing difference, the RMSD value for P9.0 and P9 ranges above 5.0 Å, well above the average (Table 2). Nevertheless, the global and even local architectural features of this region are intact, with a sharp bend from P9.0 to P9 and the formation of a tetraloop-receptor contact with P5.

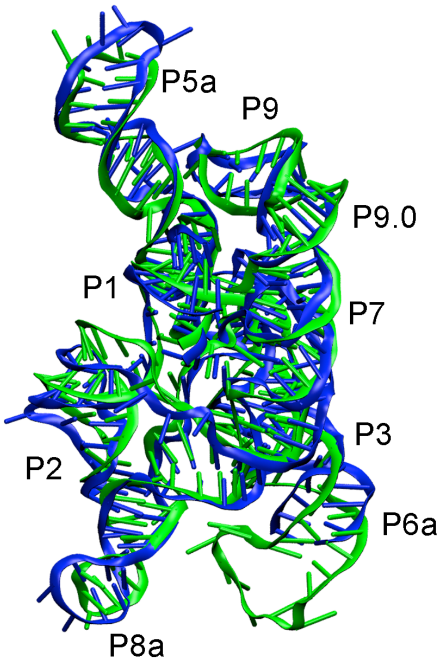

Figure 4. Model of the Azoarcus intron (blue) superimposed on the structure determined by x-ray crystallography (green). Visible helices are labeled (see Fig. 2).

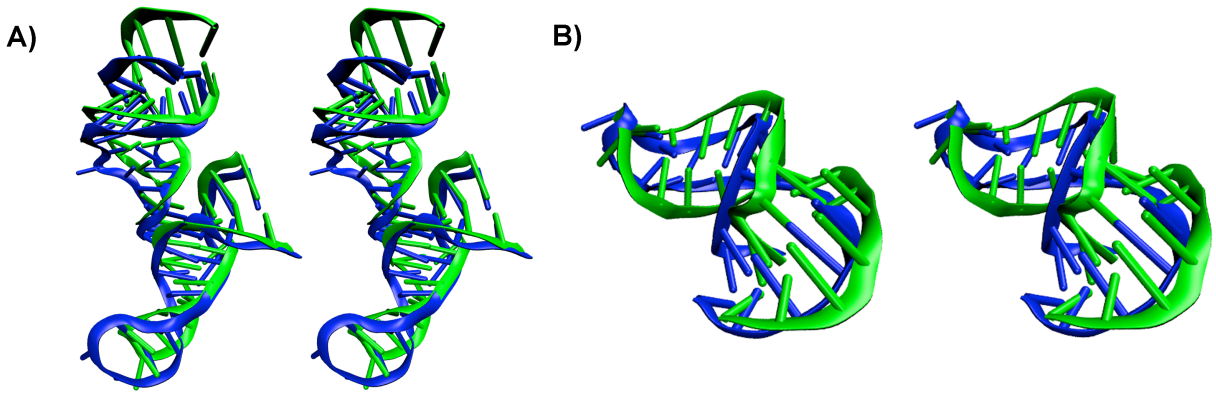

Figure 5. Regions of the Azoarcus intron model (blue) superimposed on the corresponding regions of the crystal structure (green). A) Tetraloop-receptor interaction of L2 and P8. B) P9 and P9.0. Each region is shown in wall-eyed stereoview in the same orientation as Figure 3.

Table 2

Region	Nucleotide range	RMSD (Å)
P1	10-12, 1-3 of the substrate oligonucleotide	2.55
P2	13-37	4.34
J2/3	38-40	7.20
P3	41-47, 136-143	3.45
J3/4	48-50	5.55
P4	51-56, 88-93	3.11
J4/5	57-59, 85-87	2.75
P5/5a	60-84	5.62
P6	94-96, 122-124	3.37
J6/6a	97-99, 120-121	3.22
P6a	100-119, 108-111 omitted	4.90
J6/7	125-127	3.94
P7	128-134, 173-178	4.01
P8	144-166	4.34
J8/7	167-172	3.66
P9.0	179-181, 203-205	4.90
J9/9.0	182-183, 198-202	6.00
P9	184-197	5.96
wG	206	3.05

5. Discussion

We used the RNABuilder software package to predict the structure of the *Azoarcus* group I intron by homology modeling to multiple templates. RNABuilder is an MRM method in that the templates, bases, and converged part of the model are rigidified, while forces are simultaneously treated at multiple levels of granularity. The model generated using this method displayed good agreement to the solved crystal structure of the *Azoarcus* intron, suggesting that this approach will be useful for a variety of applications in the structural biology of RNA.

RNABuilder takes advantage of several features of the SimTK core toolkit (https://simtk.org), including multibody mechanics and the Contact subsystem, which make the software well-suited for applications such as that described here. Because they permit accurate physical simulation of constrained models, these features are valuable for generating structural predictions by modeling against known RNA structures or fragments. Multibody mechanics is faster for simulations with many constraints, whereas in Cartesian mechanics constraints add computational cost. In homology modeling it is useful to constrain portions of the structure while others are built, making multibody mechanics the logical choice. For the regions that are allowed to move, the contact spheres used by RNABuilder are an economical way to approximately detect and prevent steric clashes. With these features, RNABuilder is an exceptionally efficient software package for modeling structures of large, highly constrained systems like group I introns and other structured RNAs.

In determining how to model the *Azoarcus* intron, we took steps to ensure that we were not inadvertently introducing information from the *Azoarcus* intron crystal structure. For the alignments, we only used a secondary structure model of the *Azoarcus* intron that was generated before the structure was solved (28). For regions that required decisions about how to thread them to the *Twort* intron structure, we adopted a systematic approach in which we maximized the correspondence to the *Twort* intron. In the region of P9, where the correspondences were the least clear (see Methods), this strategy resulted in successful prediction of the P9.0 base pairs but caused us to incorrectly predict two base pairs within P9. The tetraloop-receptor partners, which we modeled by threading to the

Tetrahymena intron fragment, are identical to the *Tetrahymena* intron sequences and had been predicted to share the same structure (28).

Although the agreement between the model and the *Azoarcus* crystal structure was strong throughout the molecule, it is instructive to consider the regions that gave the highest similarity and the regions that were lower. The most striking agreement was found within the core of the intron (Table 1), as might be expected because this is the most highly conserved region. Therefore, threading to even the distantly-related group I intron from *Twort* gave a good structural model for the *Azoarcus* ribozyme. The RMSD in the active site is somewhat higher, mostly because this region contains P9.0 and J9/9.0 which had higher disagreement (see below). On the other hand, the agreement was less strong in the periphery, where the conservation between the *Azoarcus* and *Twort* introns is much weaker.

Lower accuracy was also observed in regions where the peripheral architectures differ between the introns. The *Twort* intron possesses additional peripheral elements, relative to the *Azoarcus* intron, which influence the connections between structural domains. Specifically, the junction between P9 and P9.0 is part of a three-helix junction in the *Twort* intron, whereas it is a two-helix junction in the *Azoarcus* intron (see Fig. 2). Analogously, the connection between P7 and P3 is part of a complex multi-helix junction in the *Twort* intron, whereas it is a simple stacked-helix connection in the *Azoarcus* intron. In light of these substantial structural differences, it is striking that the threading approach can give as good agreement as it does in these regions. Presumably, an important factor favoring accurate modeling here is that the *Azoarcus* intron possesses a minimal set of peripheral structure elements, and it is therefore essential for the function of the intron that each peripheral element conform to the orientations and contacts found in homologous structural elements in other introns.

The success of the modeling here suggests that this approach is likely to be useful for modeling of even larger and more complex RNAs. The use of internal coordinates (in which dynamics are computed in linear time) and a force field consisting only of user-imposed interactions (avoiding long-range physical forces whose calculation scales poorly) means that significantly larger molecules can be treated with only a proportionate increase in cost. Further, some portions converge early and these can be rigidified while the rest of the threading continues, reducing computational cost. In this work the helices of azoarcus that were not threaded were formed by manually specifying base pairing contacts; by extension larger regions or even entire molecules could potentially be formed by specifying a sufficient number of such contacts. Alternatively these regions could be formed using MC-Sym, DMD, or NAST, which can build molecules in the 50-150 residue size range.

In one potential extension of this work, recent crystal structure determination and model of group II introns may prove useful for modeling of distantly related group II introns (35, 36). It may also be possible to model the ribosomal subunits as well as the much larger complete ribosomes, viral genomes (37), and spliceosomal complexes using this method. Key to the extension of our method is the ability to model larger regions of the molecule without templates and to handle such molecules without excessive computer time and memory requirements. The existing code makes substantial progress in these directions. In future work, continued improvements in speed, memory usage, and accuracy of the force fields will be useful for modeling of larger molecules and complexes.

6. Distribution

A binary distribution of RNABuilder is available for download from the RNAToolbox project on the Stanford Simbios Center's website, SimTK.org. A tutorial is available in the "Downloads" section. Simbios also provides software support and workshops.

Acknowledgements

We thank Chris Bruns, Michael Sherman, Jack Middleton, and Mark Friedrichs for substantial explanations and help with Simbody, as well as adding useful features to that code for our use. We also thank Charles Janac for help with the parameterization of the RNABuilder force field. S. Flores is supported by Simbios, the NIH Roadmap for Medical Research; Grant number: U54 GM072970 to R.B.A. This work was also supported by grants to R.R. from the National Institutes of Health (GM 070456), the Welch Foundation (F-1563), and the Norman Hackerman Advanced Research Program (003658-0242-2007).

References

1. A. Roth and R.R. Breaker, *Annu. Rev. Biochem.,* 2009, **78**, 305-334.
2. D.P. Bartel, *Cell,* 2004, **116**, 281-297.
3. R.W. Carthew and E.J. Sontheimer, *Cell,* 2009, **136**, 642-655.
4. S. Katayama, et al., *Science,* 2005, **309**, 1564-1566.
5. P. Carninci, et al., *Science,* 2005, **309**, 1559-1563.
6. T.A. Cooper, L. Wan, and G. Dreyfuss, *Cell,* 2009, **136**, 777-793.
7. A.R. Ferre-D'Amare, K. Zhou, and J.A. Doudna, *J. Mol. Biol.,* 1998, **279**, 621-631.
8. D.K. Treiber and J.R. Williamson, *Curr. Opin. Struct. Biol.,* 1999, **9**, 339-345.
9. R. Russell, *Front. Biosci.,* 2008, **13**, 1-20.
10. R. Das and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.,* 2007, **104**, 14664-14669.
11. M. Parisien and F. Major, *Nature,* 2008, **452**, 51-55.
12. F. Major, D. Gautheret, and R. Cedergren, *Proc. Natl. Acad. Sci. U. S. A.,* 1993, **90**, 9408-9412.
13. M.A. Jonikas, et al., *RNA,* 2009, **15**, 189-199.
14. C.S. Tung, S. Joseph, and K.Y. Sanbonmatsu, *Nat. Struct. Biol.,* 2002, **9**, 750-755.
15. K. Kruger, et al., *Cell,* 1982, **31**, 147-157.
16. T.R. Cech, *Annu. Rev. Biochem.,* 1990, **59**, 543-568.
17. G.J. Narlikar and D. Herschlag, *Annu. Rev. Biochem.,* 1997, **66**, 19-59.
18. S.A. Strobel and J.A. Doudna, *Trends Biochem. Sci.,* 1997, **22**, 262-266.
19. G.S. Ayton, W.G. Noid, and G.A. Voth, *Curr. Opin. Struct. Biol.,* 2007, **17**, 192-198.
20. N. Vaidehi, A. Jain, and W.A. Goddard, 3rd, *J. Phys. Chem.,* 1996, **100**, 10508-10517.
21. W.F. van Gunsteren and H.J.C. Berendsen, *Molecular Biophysics,* 1977, **34**, 1311-1327.
22. J.P. Schmidt, et al., *Proceedings of the IEEE,* 2008, **96**, 1266-1280.
23. R.E. Crosbie and W. Heyes, *Appl. Math. Modeling,* 1976, **1**, 137-140.
24. M.C. Lin, *Ph.D. thesis, University of California, Berkeley, CA,* 1993,
25. N.B. Leontis and E. Westhof, *Curr. Opin. Struct. Biol.,* 2003, **13**, 300-308.
26. S.M. Freier, et al., *Proc. Natl. Acad. Sci. U.S.A.,* 1986, **83**, 9373-9377.
27. E. Stofer, C. Chipot, and R. Lavery, *J. Am. Chem. Soc.,* 1999, **121**, 9503-9508.
28. B.L. Golden, H. Kim, and E. Chase, *Nat. Struct. Mol. Biol.,* 2005, **12**, 82-89.
29. M. Costa and F. Michel, *EMBO J.,* 1995, **14**, 1276-1285.
30. M. Costa and F. Michel, *EMBO J.,* 1997, **16**, 3289-3302.
31. J.H. Cate, et al., *Science,* 1996, **273**, 1678-1685.
32. P.L. Adams, M.R. Stahley, A.B. Kosek, J. Wang, and S.A. Strobel, *Nature,* 2004, **430**, 45-50.
33. P.L. Adams, et al., *RNA,* 2004, **10**, 1867-1887.
34. J.J. Cannone, et al., *BMC Bioinformatics,* 2002, **3**, 2-32.
35. N. Toor, K.S. Keating, S.D. Taylor, and A.M. Pyle, *Science,* 2008, **320**, 77-82.
36. L. Dai, et al., *Mol. Cell,* 2008, **30**, 472-485.
37. M.J. Roossinck, D. Sleat, and P. Palukaitis, *Microbiol. Rev.,* 1992, **56**, 265-279.