BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 24 2009, pages 3259–3266
doi:10.1093/bioinformatics/btp576

Structural bioinformatics

Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models

Magdalena A. Jonikas[1], Randall J. Radmer[1] and Russ B. Altman[1,2,*]

[1]Department of Bioengineering and [2]Department of Genetics, Stanford University, Stanford, CA 94305, USA

Received on April 23, 2009; revised on September 28, 2009; accepted on October 1, 2009

Advance Access publication October 7, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: The recent development of methods for modeling RNA 3D structures using coarse-grain approaches creates a need to bridge low- and high-resolution modeling methods. Although they contain topological information, coarse-grain models lack atomic detail, which limits their utility for some applications.

Results: We have developed a method for adding full atomic detail to coarse-grain models of RNA 3D structures. Our method [Coarse to Atomic (C2A)] uses geometries observed in known RNA crystal structures. Our method rebuilds full atomic detail from ideal coarse-grain backbones taken from crystal structures to within 1.87–3.31 Å RMSD of the full atomic crystal structure. When starting from coarse-grain models generated by the modeling tool NAST, our method builds full atomic structures that are within 1.00 Å RMSD of the starting structure. The resulting full atomic structures can be used as starting points for higher resolution modeling, thus bridging high- and low-resolution approaches to modeling RNA 3D structure.

Availability: Code for the C2A method, as well as the examples discussed in this article, are freely available at www.simtk.org/home/c2a.

Contact: russ.altman@stanford.edu

1 INTRODUCTION

Large RNA molecules perform diverse functions within cells and have complex 3D structures. For example, the 3D structures of RNA enzymes (ribozymes) allow them to catalyze RNA cleavage (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982; Stark *et al.*, 1978). Other structured functional RNA molecules include riboswitches (Nahvi *et al.*, 2002) and ribosomal RNA (Ban *et al.*, 2000; Yusupov *et al.*, 2001). Although knowing the 3D structures of these molecules is critical in understanding their functions, the protein databank contains few high-resolution RNA crystal structures. For these reasons, computational modeling of RNA 3D structures is an important goal.

Both manual and automated methods for building full atomic RNA structures have had success but continue to be challenged by significant limitations. Manual approaches have successfully modeled a number of molecules, including several group I introns (Lehnert *et al.*, 1996; Michel and Westhof, 1990) and the yeast phenylalanine tRNA (Levitt, 1969), but require expert knowledge of RNA structure. Semi-automated methods, such as MANIP (Massire and Westhof, 1998) and ERNA-3D (Tanaka *et al.*, 1998) use a fragment-based approach to build full atomic 3D RNA models. However, these methods are limited by the need for significant user interaction as well as expert knowledge.

Several automated tools model RNA structure in 3D. The MC-Fold/MC-Sym pipeline uses constraint satisfaction algorithms to build full atomic models of structures (Major *et al.*, 1993; Parisien and Major, 2008). FARNA (Das and Baker, 2007) is an automated fragment-based tool that has been used to model molecules as large as 158 nt when combined with multiplexed hydroxyl radical cleavage analysis (MOHCA; Das *et al.*, 2008).

Coarse-grain approaches to modeling RNA 3D structures are fully-automated, can model very large structures (>160 nt) and can be used in multiscale approaches for modeling large systems. These methods include YAMMP (Malhotra *et al.*, 1994), YUP (Tan *et al.*, 2006), DMD (Ding *et al.*, 2008), RNA2D3D (Martinez *et al.*, 2008) and NAST (Jonikas *et al.*, 2009). Although coarse-grain topological models of RNA molecules provide significantly more structural information than secondary structure alone, full atomic models are preferable for studying structure function relationships, and are a prerequisite for most energy-based dynamics and refinement methods.

As a result, several of these methods include tool-specific protocols for adding atomic resolution to their coarse-grain models. For example, DMD and its related web-based tool iFoldRNA (Sharma *et al.*, 2008) use a coarse-grained approach to generate predictions for structures <50 nt in size, which are then refined to atomic resolution by an unpublished reconstruction protocol. RNA2D3D starts with a nucleotide-resolution first approximation of the molecule 3D structure based on secondary structure information, and adds atomic resolution early in its protocol, using nucleotide geometries observed in crystal structures. Neither of these tools can be used on independently generated coarse-grain structures.

In some cases, users of these tools develop their own application-specific approach to adding atomic detail. A recently reported all-atomic model of Pariacoto virus included coarse-grained modeling of RNA followed by addition of atomic detail (Devkota *et al.*, 2009). The authors extended a crystal structure that contained 35% of the RNA structure by initially modeling the RNA with YAMMP, which uses a coarse-grained one-point-per-residue representation centered on phosphate atoms. To add full atomic

*To whom correspondence should be addressed.

detail to the coarse-grain structure, the authors used a fragment-based approach to match existing full atomic structures in the PDB to both base-paired and non-base-paired regions. The authors searched for compatible pieces to generate plausible structures, which they minimized and annealed. Such multiscale approaches are becoming increasingly promising, but are limited by the need for each research group to develop their own protocols for adding atomic detail to coarse-grain models.

Several methods in both RNA and protein structure prediction use knowledge-based fragment approaches, most notably FARNA and Rosetta (Simons *et al.*, 1997), these methods search for similar fragments at the sequence level to inform the structure prediction. Our approach to instantiating full atomic detail is inspired by these methods and based on the observation that fragments in RNA molecules that are similar at the coarse-grain level are often also similar at the full atomic level. Therefore, if a ribosomal RNA fragment is similar to a model fragment at the coarse-grain level, the full atomic detail from the ribosomal crystal structure may contain useful geometric information about how full atomic detail should be instantiated in the coarse-grain model.

None of the methods for adding full atomic detail to coarse-grain structures mentioned above are (i) generalized for many types of independently generated coarse-grain structures, (ii) validated on a range of structures sizes and (iii) publicly available. We present a fully automated fragment- and knowledge-based method, called Coarse to Atomic (C2A) for instantiating full atomic detail into coarse-grain structures of RNA molecules. We evaluate the full atomic structures generated by our method, and make the method freely available (along with a manual). Our method can use any atom-based coarse-grain structure template as input, and provides a full atomic, GROMACS energy-minimized structure as output. We use geometric information from RNA crystal structures to reconstruct full atomic detail. We have tested our method on both ideal coarse-grain structures for molecules ranging in size from 70 to 244 residues, as well as on coarse-grain one-point-per-residue models of tRNA generated by NAST. In addition to improving the usefulness of coarse-grain modeling methods, this tool has the potential to bridge the computationally expensive but precise full atomic modeling approach and the fast but detail-poor coarse-grain modeling approach.

2 METHODS

In brief, C2A searches within a reference full atomic structure for coarse-grain matches to fragments of a target molecule, and combines the full atomic versions of the matches to generate a full atomic structure. We then perform a minimization step to reduce any collisions and gaps. We need the following three inputs to perform this method:

- A coarse-grain template for the target molecule (e.g. a structure with one point per residues representing the C3′ or P atom) (Fig. 1A).
- A fragment definition for the target molecule (e.g. the secondary structure) (Fig. 1B).
- A reference full atomic RNA 3D structure database (e.g. ribosomal RNA structures) (Fig. 1C).

Based on this input, we generate a full atomic structure by following these steps:

(i) Use the fragment definition to break the target molecule into structural subsets (e.g. helices, loops and junctions) (Fig. 1D).

(ii) Find coarse-grain matches for each fragment in a reference full atomic RNA structure (Fig. 1E).

(iii) Combine matches to generate a full atomic structure free of major atomic collisions (Fig. 1F).

(iv) Minimize the structure to eliminate any chemically unrealistic gaps or collisions (Fig. 1G).

2.1 Defining structure fragments

For our method, we define substructures of the target molecule, which we call fragments. Although we use the secondary structure of a target molecule to define these fragments in this article, other fragment definitions that use both single- and double-stranded regions can be easily implemented as well. The fragment definition is provided by the user, and C2A parses the definition to determine the substructures. We use the secondary structure to define two types of fragments within the target molecule: single and double stranded. Helices are double-stranded base-pairing regions and may include bulges of any length. Regions between helices are single stranded and overlap by one residue in the primary sequence with adjacent helices. Fragments may not contain fewer than four residues, and more than one residue of overlap is allowed in cases where there would otherwise be fewer than four residues in a fragment. The result is a ensemble of structural coarse-grain subsets, either double or single stranded, of the target molecule (we show examples in Fig. 1D).

2.2 Searching for fragment matches

We search within the reference full atomic database of structures for coarse-grain matches to the fragments we have defined (Fig. 1E). We then use the full atomic detail of the matches to build complete full atomic structures. In this article, we use the *Thermus thermophilus* 16S ribosomal RNA solved at 3.00 Å resolution (PDB ID 1N32 chain A) as the primary database of full atomic structure. This database does not contain any structures upon which we tested the method. In our search for coarse-grain matches, we only consider the level of coarse-graining that is represented in the template structure. Thus, for NAST structures, we use template structures where each nucleotide is represented by the C3′ atom position. When searching for matches, we will only consider the C3′ positions of each residue within the reference database. However, our method allows other coarse-graining schemes such as residues represented with the position of the P atom or by more than one atom. Since using P atoms is a more common coarse-grained representation, we have provided an example that uses a P-centered coarse-grain scheme in the examples directory of the download associated with this article. For single-stranded fragments such as loops, junctions and ends, we determine the length of the fragment (in number of residues) and the distance between the defining points of the first and last residues in the template structure (in Angstroms). We search through the reference database for continuous runs of residues of the same length as the fragment. Of these, we keep the ones that have distances between the first and last residue defining atoms (in our case C3′) within a certain user-specified error (typically 10%) of the distance observed in the template fragment. This results in an ensemble of matches within the full atomic structure to the coarse-grain template fragment. We then calculate the coarse-grain RMSD (Root Mean Square Deviation) of each potential match to the coarse-grain template fragment and rank the matches. We keep the 10 best matches by RMSD for each fragment as part of the working set for assembling a full atomic structure. For double-stranded fragments such as helices, we first find all the possible matches for each single-stranded element individually. From each possible combination of matches, we keep those that are within a certain error (we use 10%) of the distances observed in the template fragment between the 5′ and 3′ ends of each fragment. As with the single-stranded fragments, we calculate the coarse-grain RMSD to

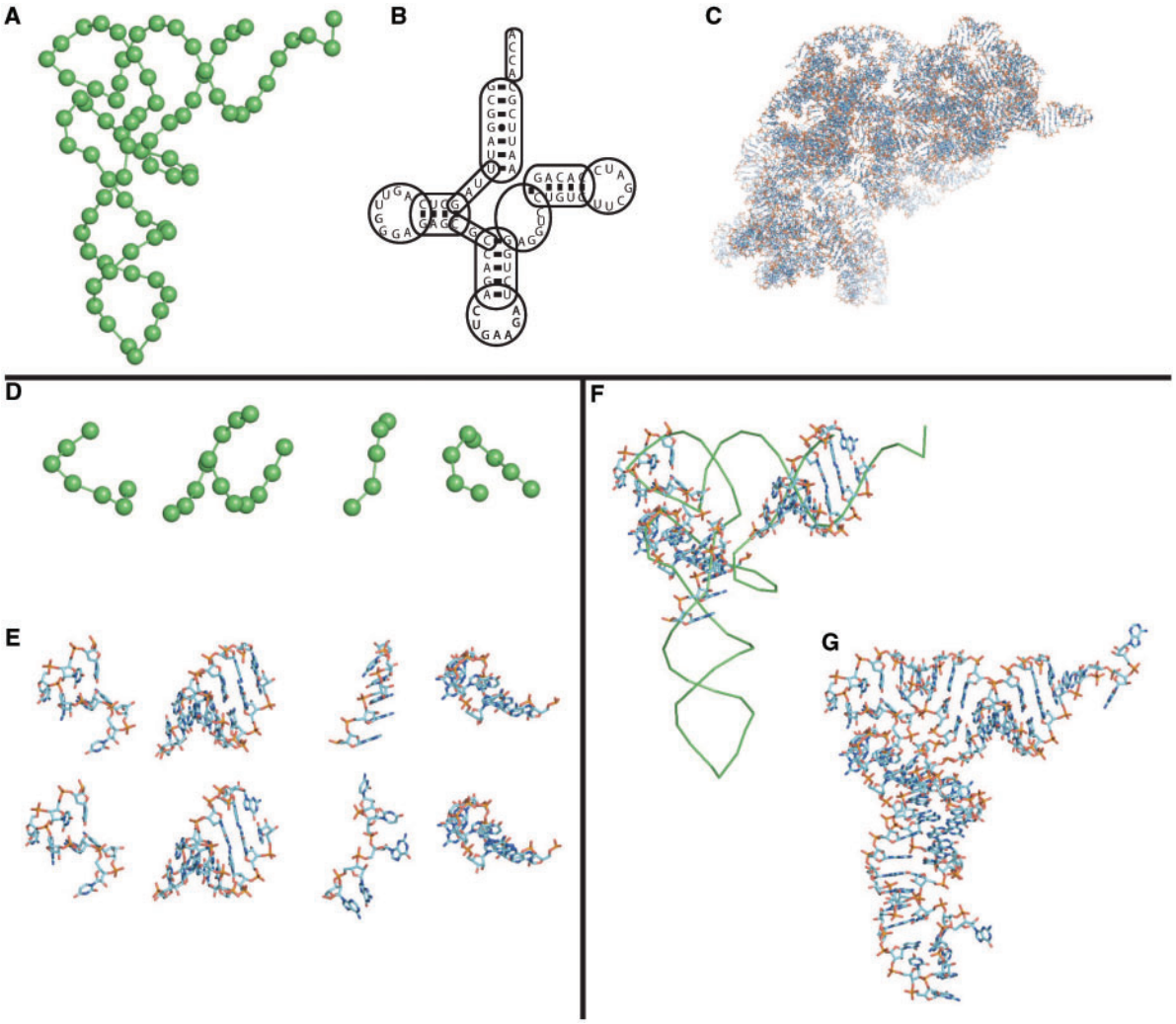

Fig. 1. C2A method overview. The C2A method uses three pieces of information as input: (**A**) a coarse-grain template for the target molecule; (**B**) a fragment definition for the target molecule, such as the secondary structure; and (**C**) a reference database of full atomic RNA structures. (**D**) In Step 1, the coarse-grain template is divided into structural subsets based on the fragment definition, here we show as an example a double- and a single-stranded fragment. (**E**) Step 2 searches for coarse-grain matches to each fragment in the reference structure and extracts the full atomic detail of each match, we show here the full atomic detail for two matches found in ribosomal RNA for each of the two example fragments. (**F**) Step 3 searches for combinations of matches free of major collisions, we show here how matches for the two example fragments align to the coarse-grain template. (**G**) Finally, Step 4 minimizes the resulting full atomic structure to remove unrealistic gaps and collisions.

the coarse-grain template fragment, rank the matches and keep the 10 best for the working set.

2.3 Processing fragment matches

Since we do not consider the residue sequence in our search for matches, many of the matches will have incorrect base atoms. To remedy this, we replace the incorrect nucleoside base atoms with the correct ones while maintaining the orientation of the base plane. We keep the sugar group and phosphate group atoms in the same positions, and replace the base atoms with the correct geometry. We use three atoms to define the plane and orientation of the original base (N9, N1 and C5 for Adenine and Guanine, and N1, N3 and C5 for Cytosine and Uracil). We insert the correct base atoms into the appropriate plane and orientation, resulting in coordinates for the correct residue. This results in fragment matches with the correct nucleotide sequence while maintaining the geometry of the match found in the reference structure.

2.4 Assembling full atomic structures

We use a Metropolis Monte Carlo approach to assemble a full atomic structure using the coarse-grain template and a working set of best matches for each fragment. We start by randomly selecting one match for each fragment from the working set and aligning the matches to the coarse-grain template. Since we defined single-stranded fragments to overlap by one residue with double-stranded fragments, each of these junctions will have two sets of coordinates for that residue. In these cases, we use only the coordinates from the double-stranded fragment. In each step, we search for pairs of atoms within the full atomic structure that are closer than a cutoff distance (user-specified, 1.0 Å here) and check the fragments containing these pairs for collisions. If we observe any collisions, we randomly select one of

the colliding fragments to be replaced with another randomly selected match. We accept the replacement with the probability shown in Equation (1), where N_{old} is the number of collisions in the old state and N_{new} is the number of collisions in the new state.

$$\begin{aligned} &\text{if} \quad N_{new} < N_{old} \\ &\text{then} \quad P = 1.0 \\ &\text{else} \quad P = 0.5 \times \exp\left[-\frac{N_{new} - N_{old}}{N_{old}}\right] \end{aligned} \tag{1}$$

We continue this attempt for a user-specified number of steps, or until we find a combination of matches with no collisions, whichever occurs first. If no such combination is found within a user-defined number of steps and attempts, the output will be the full atomic structure with least number of collisions observed over the course of this search. In this article, we make five attempts of 500 steps each to generate a full atomic structure free of major collisions.

2.5 Minimizing full atomic structures

The structures resulting from our assembly protocol are likely to contain both gaps (unrealistically long covalent bonds) and collisions (unrealistically short distances between any atoms), and which may prevent the structures from being used in full atomic studies. In particular, the regions at junctions between fragments may contain significant gaps between atoms that should be bonded because the coordinates for those atoms came from different matches. To remedy these two structural issues, we minimize the full atomic structures with GROMACS using steepest descent minimization. We include the scripts we used for running GROMACS, including parameter files, on our project's web site.

2.6 Evaluating full atomic structures

We evaluated the effect of minimization on junction bonds by calculating the lengths of covalent bonds between fragments both before and after minimization. We compared these values with those of non-junction bonds. We also calculated the lengths of these bonds observed in the relevant crystal structures.

We submitted our minimized structures to the MolProbity tool, which calculates a Clashscore and assigns a percentile to the structure (Davis *et al.*, 2007; Lovell *et al.*, 2003). We also submitted our structures to the RCSB ADIT structure validation tool which outputs a full report on the structure including RMSD values for covalent bonds and angles relative to standard values for nucleotides (Clowney *et al.*, 1996; Gelbin *et al.*, 1996). We made the full reports available on our project web site under the documents section.

For all minimized full atomic structures, we calculated the RMSD value relative to the known crystal structure. We also calculated RMSD values for helical (double-stranded) and non-helical (single-stranded) fragments separately. We also applied the recently developed metric of interaction network fidelity (INF) to all of our structures, which considers base–base-stacking and base–base-pairing interactions (Parisien *et al.*, 2009). We calculated the INF for pairings alone, as well as pairings and stackings combined, which is a much stricter measure. INF values range from 0.00 (worst) to 1.00 (best).

2.7 Validation using ideal backbones from crystal structures

To validate our approach, we applied it to seven crystal structures of RNA molecules ranging in size from 70 to 244 residues:

- Yeast ai5g group II intron ('Ai5gamma', PDB ID 1KXK) 70 residues (Zhang and Doudna, 2002).
- Yeast phenylalanine tRNA ('tRNA', PDB ID 6TNA) 76 residues (Sussman *et al.*, 1978).
- Aminoacyl-tRNA synthetase ribozyme ('flexizyme', PDB ID 3CUL) 92 residues (Xiao *et al.*, 2008).
- P4-P6 RNA ribozyme domain ('P4-P6', PDB ID 1GID) 158 residues (Cate *et al.*, 1996).
- Azoarcus group I intron ('Azoarcus', PDB ID 1ZZN) 195 residues (Stahley and Strobel, 2005).
- Twort ribozyme ('Twort', PDB ID 1Y0Q) 244 residues (Golden *et al.*, 2005).

We stripped each crystal structure of atomic detail, leaving only the position of the C3′ atom for each residue. We defined fragments using the known secondary structure of each molecule and the rules described above. The database did not contain any of these structures. We searched for matches to each fragment and created a working set by keeping the 10 best matches by RMSD at the coarse-grain level. We searched for 10 combinations of matches free of major collisions and minimized the full atomic structures. For the two structures we modeled that were missing the residues (the *Twort* and *Azoarcus* ribozymes), we used the complete coarse-grain structures generated using the NAST tool by Jonikas *et al.* as the template input.

2.8 Building full atomic models of NAST tRNA topologies

We applied the C2A method to the problem of building full atomic structures based on coarse-grain models built by the tool NAST. For each of the three topologies generated by NAST, we used five models as template starting structures and generated 10 full atomic structures which we then minimized.

3 RESULTS

3.1 Recovering full atomic detail from ideal coarse-grain templates

The full atomic RMSD values of full atomic models built from coarse-grain crystal structure templates ranged from 1.87 to 3.31 Å, we give the average values in Table 1. The average RMSD for the 60 full atomic structures we generated from ideal backbones was 2.75 ± 0.37 Å. We report INF scores for all structures in Table 2. We calculated junction bond-distances before and after minimization and report the values in Table 3. We report MolProbity Clashschore and percentiles, as well as RMSD values for covalent bonds and angles as evaluated by the RCSB ADIT tool in Table 4. We compare

Table 1. RMSD of minimized full atomic structures

Molecule ID	Overall RMSD (Å)	Helical fragments RMSD (Å)	Non-helical fragments RMSD (Å)
Validation structures (10 models)			
1KXK	2.13 ± 0.21	1.36 ± 0.44	2.41 ± 1.39
6TNA	2.81 ± 0.11	1.10 ± 0.09	3.02 ± 1.27
3CUL	3.06 ± 0.18	1.77 ± 0.82	2.48 ± 1.90
1GID	3.16 ± 0.08	2.03 ± 1.22	2.75 ± 1.45
1ZZN	2.79 ± 0.16	2.10 ± 0.74	2.56 ± 1.42
1Y0Q	2.76 ± 0.07	1.92 ± 0.68	2.50 ± 1.54
NAST tRNA models (50 models)			
Model A	8.39 ± 0.27	2.65 ± 0.71	4.13 ± 1.91
Model B	13.30 ± 0.31	3.08 ± 1.28	3.82 ± 1.51
Model C	15.99 ± 0.76	3.06 ± 1.13	4.19 ± 1.95

We report both overall RMSD values and separate values for helical and non-helical fragments.

Table 2. Averages and ranges for INF for base pairs alone and base pairs with stacking

Molecule ID	INF on pairings and stacking		INF on pairings alone	
	Average	Range	Average	Range
1KXK	0.51 ± 0.10	$0.44 - 0.78$	0.83 ± 0.03	$0.77 - 0.89$
6TNA	0.69 ± 0.01	$0.66 - 0.71$	0.69 ± 0.04	$0.59 - 0.73$
3CUL	0.67 ± 0.02	$0.64 - 0.70$	0.65 ± 0.04	$0.57 - 0.69$
1GID	0.61 ± 0.03	$0.56 - 0.64$	0.51 ± 0.03	$0.46 - 0.55$
1ZZN	0.58 ± 0.02	$0.55 - 0.60$	0.67 ± 0.02	$0.64 - 0.70$
1Y0Q	0.61 ± 0.01	$0.58 - 0.63$	0.54 ± 0.02	$0.52 - 0.57$
6TNA-A	0.46 ± 0.04	$0.38 - 0.55$	0.35 ± 0.07	$0.18 - 0.51$
6TNA-B	0.41 ± 0.05	$0.31 - 0.51$	0.27 ± 0.09	$0.07 - 0.46$
6TNA-C	0.41 ± 0.04	$0.34 - 0.49$	0.26 ± 0.08	$0.11 - 0.45$

INF scores range from 0.0 (worst) to 1.0 (best).

Table 3. Covalent bond lengths before and after minimization for junctions and non-junctions

Molecule ID	Non-junction bonds (Å)		Junction bonds (Å)	
	Range	Average	Range	Average
1KXK				
Crystal	$1.58 - 1.61$	1.60 ± 1.60	$1.59 - 1.63$	1.60 ± 0.01
Pre-min	$1.59 - 1.63$	1.61 ± 1.61	$1.12 - 6.24$	3.29 ± 1.22
Post-min	$1.54 - 1.65$	1.59 ± 1.59	$1.56 - 1.75$	1.64 ± 0.04
6TNA				
Crystal	$1.54 - 1.64$	1.60 ± 1.60	$1.58 - 1.62$	1.60 ± 0.01
Pre-min	$1.59 - 1.62$	1.60 ± 1.60	$1.01 - 6.49$	2.49 ± 1.21
Post-min	$1.53 - 1.65$	1.59 ± 1.59	$1.51 - 1.78$	1.62 ± 0.05
3CUL				
Crystal	$1.48 - 1.62$	1.60 ± 1.60	$1.59 - 1.62$	1.61 ± 0.01
Pre-min	$1.59 - 1.62$	1.61 ± 1.61	$1.14 - 8.26$	4.07 ± 1.69
Post-min	$1.55 - 1.65$	1.59 ± 1.59	$1.56 - 1.86$	1.67 ± 0.07
1GID				
Crystal	$1.57 - 1.64$	1.60 ± 1.60	$1.58 - 1.62$	1.60 ± 0.01
Pre-min	$1.59 - 1.64$	1.61 ± 1.61	$1.04 - 12.72$	3.82 ± 2.50
Post-min	$1.49 - 1.70$	1.59 ± 1.59	$1.50 - 2.90$	1.67 ± 0.13
1ZZN				
Crystal	$1.57 - 1.63$	1.61 ± 1.61	$1.57 - 1.64$	1.61 ± 0.01
Pre-min	$1.59 - 1.63$	1.60 ± 1.60	$1.00 - 13.36$	3.64 ± 1.67
Post-min	$1.42 - 2.09$	1.59 ± 1.59	$1.44 - 2.47$	1.67 ± 0.11
1Y0Q				
Crystal	$1.58 - 1.65$	1.60 ± 1.60	$1.57 - 1.64$	1.60 ± 0.01
Pre-min	$1.59 - 1.63$	1.61 ± 1.61	$0.55 - 10.96$	3.53 ± 1.88
Post-min	$1.46 - 1.87$	1.59 ± 1.59	$1.48 - 2.13$	1.66 ± 0.07

We report ranges as well as average values. We also report the bond lengths in the relevant crystal structures for the same bonds.

these values for our models to those calculated on the relevant crystal structures. We show a sample full atomic tRNA molecule built from the ideal backbone template in Figure 2B and show the full atomic crystal structure for comparison (Fig. 2A). The full atomic models we generated for the *Azoarcus* and *Twort* molecules contain full atomic detail for the loop regions that are missing in the crystal structure. We have posted these structures and all code necessary to reproduce these examples on our project web site.

3.2 Building full atomic models from coarse-grain NAST models of tRNA

We show sample full atomic models built from the three NAST topology predictions of tRNA in Figure 2C–E. The full atomic structures that we generated had full atomic RMSD values within 1 Å of the coarse-grain RMSD of their templates. We report RMSD values to the tRNA crystal structure in Table 1 along with separate values for helical versus non-helical fragments. We also report INF scores for pairings alone, as well as pairings and stackings in Table 2. We report MolProbity Clashscores and RCBS ADIT RMSD values for covalent bonds and angles in Table 5. The fragment combination search step succeeded in 50, 24 and 20 of the 50 attempts for each model A, B, and C, respectively. For model B, two of the five templates succeeded in 10 of the 10 combination search attempts, one succeeded in 4 of the 10 attempts, and two succeeded in 0 of the 10 attempts. For model C, two templates succeeded in 10 of the 10 attempts and three succeeded in 0 of the 10 attempts.

3.3 Computational resources

We used a 2.2 GHz CPU to run the C2A method. We analyzed the CPU load for the two parts of C2A code: creating a working library of fragment matches (which only needs to be performed once per template structure) and combining fragments into full atomic structures (which we performed 10 times for each template structures).

There are two contributing factors to the time needed to generate the working library of fragment matches: the number and types of fragments, and the size of the reference full atomic database. It took 95 s to load the reference database [*Thermus thermophilus* 16S ribosomal RNA solved at 3.00 Å resolution (PDB ID 1N32 chain A)]. It took between 3 s and 20 s to find matches for single-stranded fragments and between 40 s and 440 s to find matches for double-stranded fragments. In total, the first part of C2A took 1226, 1550, 1034, 1297, 1836, 2741 and 3847 s to find matches to all fragments for the Ai5gamma, tRNA, flexizyme, SPR19, P4-P6, *Azoarcus* and *Twort* molecules, respectively.

When using ideal backbones as the input for the C2A method, finding combinations of matches free of major collisions took an average of 47 ± 30 s, with the maximum being 117 s. For NAST tRNA models A, B and C, finding matches took average times of 1970, 1956 and 2206 s, respectively.

The code and examples we have made available are python scripts that have been tested on both unix and Mac OSX platforms.

4 DISCUSSION

We have presented a method for instantiating full atomic detail in coarse-grain RNA structure backbones using geometry knowledge from a reference database of one or more full atomic RNA crystal structures, in this case the *Thermus thermophilus* 16S ribosome. We assume that if a fragment matches well to pieces of known RNA structure at the coarse-grain level, then the full atomic geometry of the match is a good predictor of the fragment's full atomic detail. In this article, we used only data from one ribosomal RNA as a

Table 4. Evaluation of minimized full atomic structures with MolProbity and the RCSB ADIT tool

Molecule ID	MolProbity						RCSB ADIT			
	Clashscore (goal = 0)			Percentile ($N = 1784$)			Covalent bonds		Covalent angles	
	Crystal structure	Best model	Average (10 models)	Crystal structure	Best model	Average (10 models)	RMSD (Å)		RMSD (degrees)	
							Crystal	Models	Crystal	Models
1KXK	14.03	3.33	8.54 ± 5.54	54	97	78.30 ± 22.53	0.006	0.017	0.9	2.8
6TNA	26.46	4.92	9.11 ± 4.54	19	94	76.60 ± 19.10	0.079	0.017	3.4	2.7
3CUL	7.67	6.11	8.56 ± 2.30	83	90	79.10 ± 10.81	0.007	0.020	1.0	3.2
1GID	35.78	13.31	23.21 ± 14.67	10	57	35.80 ± 18.36	0.009	0.031	0.9	3.4
1ZZN	32.67	15.29	29.46 ± 9.82	13	48	21.11 ± 16.00	0.008	0.032	1.1	3.6
1Y0Q	40.47	15.9	40.82 ± 17.72	8	46	13.60 ± 15.03	0.007	0.036	1.0	3.8

We give MolProbity Clashscores and percentiles for validation structures. Clashscore is the number of serious steric overlaps (>0.4 Å) per 1000 atoms. ADIT reports the RMSD values for covalent bonds and angles calculated relative to standard values for nucleotide units.

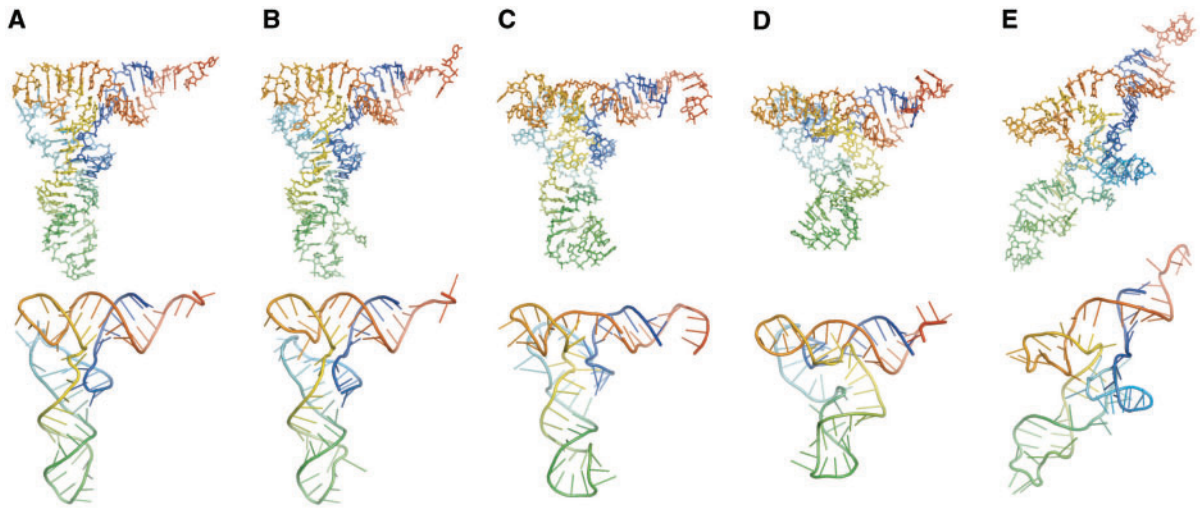

Fig. 2. Images of C2A full atomic modeling results. We show both a full atomic stick and a cartoon representation for each of the following tRNA structures: the full atomic crystal structure (PDB ID 6TNA) (**A**), a full atomic model generated by C2A using the ideal backbone from the crystal structure (**B**), a full atomic model based on the NAST tRNA model A (**C**), model B (**D**), and model C (**E**).

reference structure, through which we search for matches to coarse-grain templates. We did not include any geometric data from our template structures in the reference full atomic structure.

To validate our method, we applied it to ideal backbones taken from the crystal structures of seven RNA molecules. For each structure, we kept only the C3′ position of each residue, creating ideal backbone structures with one point per residue and recovered the full atomic detail at an average resolution <3.0 Å. We were also able to generate full atomic structures of two large RNA molecules that were missing residues in the crystal structure solution (the *Azoarcus* and *Twort* ribozymes) by using the NAST generated complete coarse-grain templates. We were able to minimize the full atomic structures with GROMACS, and remove significant clashes and gaps. The resulting minimized full-atomic structures compare well with the known crystal structures in terms of clashes and gaps. Users with expertise in structure minimization and refinement can use these structures as starting points for further studies.

For the two ribozymes that are missing loops in the crystal structure (*Azoracus* and *Twort*), we used the coarse-grain loops modeled by NAST as part of the input template and modeled full atomic detail for the entire molecule, resulting in complete full atomic structures. These two examples show that the combination of NAST and C2A can be used to complete crystal structures that are missing pieces of the molecule. However, the quality of the modeling of the missing pieces depends entirely on the modeling choices by the user when building the coarse-grain template. In our case, we used the geometry of existing loops to build coarse-grain models of the missing loops.

We then applied our method to fill in atomic resolution in NAST coarse-grain models of the tRNA molecule. For each of the three coarse-grain topology models generated by NAST, we used five representative structures and applied our method to generate full atomic versions of the models. As an aside, we noticed that the amount of time necessary to find a combination of fragments

Table 5. MolProbity Clashscores for full atomic tRNA models built from NAST coarse-grained models

Molecule ID	MolProbity				RCSB ADIT	
	Clashscore (goal = 0)		Percentile ($N = 1784$, 0–9999 Å)		Covalent bonds	Covalent angles
	Best model	Average	Best model	Average	RMSD (Å)	RMSD (degrees)
Model A: all structures	1.23	6.93 ± 5.03	99	84.32 ± 16.67	0.022	3.2
Structures A-0	3.08	7.14 ± 3.36	98	84.2 ± 14.54	0.021	3.2
Structures A-1	1.23	9.29 ± 8.38	99	78.1 ± 24.59	0.024	3.4
Structures A-2	5.54	8.80 ± 2.51	92	77.4 ± 11.89	0.022	3.3
Structures A-3	1.23	6.52 ± 4.34	99	84.6 ± 16.32	0.022	3.2
Structures A-4	1.23	2.89 ± 1.56	99	97.3 ± 2.50	0.020	3.1
Model B: all structures	7.38	37.77 ± 26.96	85	25.16 ± 24.34	0.041	4.0
Structures B-0	7.38	13.29 ± 4.18	85	57.7 ± 18.66	0.022	3.3
Structures B-1	13.54	40.41 ± 23.72	56	18.78 ± 20.09	0.039	4.0
Structures B-2	24.00	62.83 ± 34.44	23	8.2 ± 10.53	0.050	4.5
Structures B-3	14.77	37.17 ± 17.14	50	17.4 ± 17.88	0.046	3.9
Structures B-4	14.77	35.39 ± 22.44	50	23.1 ± 20.36	0.044	4.0
Model C: all structures	3.08	19.04 ± 13.78	98	48.42 ± 28.22	0.029	3.6
Structures C-0	9.23	18.34 ± 9.96	75	46.4 ± 24.17	0.033	3.7
Structures C-1	15.38	22.65 ± 4.36	48	27.8 ± 10.41	0.023	3.5
Structures C-2	3.08	12.19 ± 7.56	98	64.5 ± 29.39	0.025	3.4
Structures C-3	6.15	11.57 ± 4.39	90	65.8 ± 19.44	0.026	3.6
Structures C-4	5.54	30.46 ± 23.76	92	37.6 ± 34.25	0.035	3.8

For each of three groups of coarse-grained models (A–C), we selected five models as templates for building 10 full atomic structures. We list statistics for each of the three groups, as well as for each coarse-grain template we used.

free of major collisions is an indicator of the quality of the topology. Additionally, all five of the templates used for model A successfully generated full atomic structures, while only three and two templates for models B and C, respectively, were able to do so. We were able to minimize the full atomic structures we built using GROMACS which corrected any gaps or collisions. Until this last minimization step, the method is entirely geometric and does not consider any chemical opportunities such as hydrogen bonds. Through this example, we show that coarse-grain topology models generated from limited structural information (in this case, only primary, secondary and limited tertiary structure data) can be used as starting points for building full atomic structures. These full atomic structures can be used as starting points for further refinement using physics-based approaches, or more detailed knowledge-based approaches.

The full atomic structures we generated for tRNA, both from the crystal structure template and the NAST model template, did not recover any tertiary contact base pairings with adequate detail. Recovering these interactions is an important goal of model RNA 3D structure and remains a significant challenge. Our need to avoid serious clashes during the assembly step selects against the close interactions needed to recover these non-helical base pairings. However, once a full-atomic structure is built from a coarse-grain template, knowledge of tertiary interactions and finer-grained dynamics tools can be used to recover these interactions. Preliminary results of using the RNAbuilder tool (S.C.Flores *et al*., submitted for publication) to constrain known tertiary contacts in our full atomic models show improved INF scores. RNAbuilder uses a full atomic model as a rigid template onto which it threads another full atomic molecule which attempts to satisfy both the geometry of the template and the specified tertiary contacts. This method maintains most of the original geometry which additionally constrains tertiary contacts that were not previously present.

The C2A method for instantiating full atomic detail into coarse-grain models is limited by the quality of the template structure and information in the reference full atomic structure. Although the process of instantiating atomic detail does not decrease the quality of the model, it does not improve it either. Our method is also limited by the geometric information contained in the chosen reference database. C2A cannot predict entirely new fragment geometries, and assumes that known structures are reasonable sources for data for the structures we are attempting to predict. Although we used only one ribosomal RNA structure, users of C2A can include different or more crystal structures as reference full atomic structural information.

Full atomic structures generated by C2A inherit the limitations and problems associated with their coarse-grain templates. Additionally, there is no guarantee of convergence on a combination of full atomic fragments free of major collisions. However, we noticed a trend that a lack of convergence tended to correlate with a coarse-grain template model of poor quality.

Our code works with any atom-based coarse-graining scheme. The results we presented in the article are for coarse-grain template structures with a one-point-per-residues representation centered on the C3′ atom. We have also tested our code on P, P-C3′ and P-C2-C3′ coarse-grain representations (not presented in this article). We observe that the more points representing a residue, the better the resulting full atomic structures.

The method we have proposed connects the coarse-grain and full atomic approaches for modeling RNA 3D structures. Coarse-grain approaches are generally faster and able to handle large molecules but lack in accuracy and detail. Full atomic approaches are limited computationally, but provide accurate and detailed results. Being able to move back and forth between these two approaches has not been possible because no published, validated and publicly available method existed for building full atomic detail into coarse-grain models. Our method will allow models generated by coarse-grain methods to be refined using full atomic tools. Using full atomic models in coarse-grain methods simply requires removal of full atomic detail to the desired coarse-grain level.

We make all code and examples from this article, along with instructions, available for free on our project web site: `www.simtk.org/home/c2a` under the Documents link. Detailed instructions for running C2A are including in the NAST/C2A manual which is available at the same address.

ACKNOWLEDGEMENTS

We thank Alain Laederach, Daniel Herschlag, Rhiju Das and Samuel Flores for helpful discussions. We are grateful to Christopher Bruns for developing the GROMACS interface Zephyr (available free at https://simtk.org/home/zephyr). We also thank Marc Parisien for providing us with code to calculate the INF scores of our structures.

Funding: NIH Roadmap for Medical Research (grant U54 GM072970); National Institutes of Health (grant P01-GM66275). National Library of Medicine Training (grant LM-07033 to M.A.J.). NIH Biotechnology Training (grant 5 T32GM008412-15 to M.A.J.).

Conflict of Interest: none declared.

REFERENCES

Ban,N. *et al*. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 a resolution. *Science*, **289**, 905–920.

Cate,J.H. *et al*. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678–1685.

Clowney,L. *et al*. (1996) Geometric parameters in nucleic acids: nitrogenous bases. *J. Am. Chem. Soc.*, **118**, 509–518.

Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.

Das,R. *et al*. (2008) Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc. Natl Acad. Sci. USA*, **105**, 4144–4149.

Davis,I.W. *et al*. (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35** (Suppl. 2), W375–W383.

Devkota,B. *et al*. (2009) Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. *Biopolymers*, **91**, 530–538.

Ding,F. *et al*. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.

Gelbin,A. *et al*. (1996) Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Am. Chem. Soc.*, **118**, 519–529.

Golden,B.L. *et al*. (2005) Crystal structure of a phage twort group I ribozyme-product complex. *Nat. Struct. Mol. Biol.*, **12**, 82–89.

Guerrier-Takada,C. *et al*. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**(Pt 2), 849–857.

Jonikas,M.A. *et al*. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.

Kruger,K. *et al*. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, **31**, 147–57.

Lehnert,V. *et al*. (1996) New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the tetrahymena thermophila ribozyme. *Chem. Biol.*, **3**, 993–1009.

Levitt,M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759–763.

Lovell,S.C. *et al*. (2003) Structure validation by calpha geometry: phi, psi and cbeta deviation. *Proteins: Structure, Function, and Genetics*, **50**, 437–450.

Major,F. *et al*. (1993) Reproducing the three-dimensional structure of a TRNA molecule from structural constraints. *Proc. Natl Acad. Sci. USA*, **90**, 9408–9412.

Malhotra,A. *et al*. (1994) Modeling large RNAS and ribonucleoprotein particles using molecular mechanics techniques. *Biophys. J.*, **66**, 1777–1795.

Martinez,H.M. *et al*. (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J. Biomol. Struct. Dyn.*, **25**, 669–683.

Massire,C. and Westhof,E. (1998) MANIP: an interactive tool for modelling RNA. *J. Mol. Graph Model*, **16**, 197–205, 255–257.

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.

Nahvi,A. *et al*. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043.

Parisien,M. and Major,F. (2008). The MC-fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.

Parisien,M. *et al*. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.

Sharma,S. *et al*. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.

Simons,K.T. *et al*. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.

Stahley,M.R. and Strobel,S.A. (2005) Structural evidence for a two-metal-ion mechanism of group I intron splicing. *Science*, **309**, 1587–1590.

Stark,B.C. *et al*. (1978) Ribonuclease P: an enzyme with an essential RNA component. *Proc. Natl Acad. Sci. USA*, **75**, 3717–3721.

Sussman,J.L. *et al*. (1978) Crystal structure of yeast phenylalanine transfer RNA. i. crystallographic refinement. *J. Mol. Biol.*, **123**, 607–630.

Tanaka,I. *et al*. (1998) Matching the crystallographic structure of ribosomal protein s7 to a three-dimensional model of the 16s ribosomal RNA. *RNA*, **4**, 542–550.

Tan,R.K.Z. *et al*. (2006) YUP: a molecular simulation program for coarse-grained and multiscaled models. *J. Chem. Theory Comput.*, **2**, 529–540.

Xiao,H. *et al*. (2008) Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme. *Nature*, **454**, 358–361.

Yusupov,M.M. *et al*. (2001) Crystal structure of the ribosome at 5.5 a resolution. *Science*, **292**, 883–896.

Zhang,L. and Doudna,J.A. (2002) Structural insights into group II intron catalysis and branch-site selection. *Science*, **295**, 2084–2088.