Review Article

Solving novel RNA structures using only secondary structural fragments

Michael P. Robertson [a,b,c], Young-In Chi [d], William G. Scott [a,*]

[a] Department of Chemistry and Biochemistry and The Center for the Molecular Biology of RNA, University of California, Santa Cruz, CA 95064, USA
[b] Department of Chemistry, The Scripps Research Institute, La Jolla, CA 92037, USA
[c] Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
[d] Center for Structural Biology, Department of Molecular and Cellular Biochemistry, University of Kentucky, 741 S. Limestone, Lexington, KY 40536, USA

A R T I C L E   I N F O

A B S T R A C T

The crystallographic phase problem is the primary bottleneck encountered when attempting to solve macromolecular structures for which no close crystallographic structural homologues are known. Typically, isomorphous "heavy-atom" replacement and/or anomalous dispersion methods must be used in such cases to obtain experimentally-determined phases. Even three-dimensional NMR structures of the same macromolecule are often not sufficient to solve the crystallographic phase problem. RNA crystal structures present additional challenges due to greater difficulty in obtaining suitable heavy-atom derivatives. We present a unique approach to solve the phase problem for novel RNA crystal structures that has enjoyed a reasonable degree of success. This approach involves modeling only those portions of the RNA sequence whose structure can be predicted readily, i.e., the individual A-form helical regions and well-known stem-loop sub-structures. We have found that no prior knowledge of how the helices and other structural elements are arranged with respect to one another in three-dimensional space, or in some cases, even the sequence, is required to obtain a useable solution to the phase problem, using simultaneous molecular replacement of a set of generic helical RNA fragments.

© 2010 Elsevier Inc. All rights reserved.

1. Background

To solve a novel macromolecular crystallographic structure, the phase problem must be solved, typically using isomorphous replacement of heavy atoms or anomalous scattering [1]. If a homologous crystal structure exists, it may be possible to use that structure for molecular replacement [2], provided that the homologous search model has an r.m.s.d. less than 1.5 Å; even homologous NMR structures often are not sufficiently similar for molecular replacement to succeed, permitting crystallographers to speculate that NMR might really stand for "Not for Molecular Replacement" [3,4].

Novel RNA crystal structures are especially challenging, given the comparative lack of homologues (in contrast to the many thousands of protein structures available in the protein data base), as well as their limited ability to form useful heavy-atom derivatives [5].

We have found, counter-intuitively, that prior knowledge of how an RNA molecule folds in three dimensions is not required for successful solution of the crystallographic phase problem via molecular replacement, provided ideal models of a subset of helical fragments corresponding to the known secondary structure of a crystallized RNA can be generated and used as a set of search models for molecular replacement. This approach has been used to solve structures of the L1 ribozyme ligase [6], satellite tobacco ringspot virus hammerhead ribozyme [7], riboswitches [8,9], and other RNA structures; a detailed description of the solution of the L1 ligase ribozyme structure has appeared elsewhere [10]. Here we focus upon the generalized method for solving RNA structures using helical fragments, and several improvements that have been incorporated since the ligase ribozyme structure was solved.

Most structured RNAs consist mainly of a set of helical elements and connecting loops. Within these helical elements, most base-pairs are either standard Watson–Crick pairs, or are variants that do not grossly perturb the secondary structure. Because of this, it is possible to predict RNA secondary structure with a reasonably high degree of confidence using standard computational techniques such as those incorporated in freely available software such as MFOLD [11] and ViennaRNA [12], and these secondary structural predictions are readily testable using standard nucleic acid biochemical probing techniques. Hence, by the time one is crystallizing an RNA molecule, its secondary structure most likely has already been established. Many RNAs also incorporate other well-known structural elements, such as the GNRA tetraloop. Hence reasonably accurate models of individual secondary structural elements of a complex RNA molecule can be obtained simply from modeling standard A-form RNA helices. We have found the

---

* Corresponding author. Fax: +1 831 459 3139.
E-mail address: wgscott@chemistry.ucsc.edu (W.G. Scott).

program *COOT* (Crystallographic Object-Oriented Toolkit) [13,14] to be the most efficient way to generate idealized model A-form RNA fragments, using the menu item "calculate > other modelling tools > ideal DNA/RNA". This instantly generates an ideal A-form RNA helix for any given sequence. If required, non-helical structures such as GNRA tetraloops may be obtained from the PDB and can be grafted onto a modeled helix. Typically, we start with up to four independent helical elements, in four separately named PDB files, even if the RNA represented by these fragments is less than 1/2 of the total RNA in the crystallographic asymmetric unit. In fact, this sort of "under-sampling" often improves the molecular replacement solution [6].

PHASER is an automated molecular replacement program [15] that is particularly well-suited to using several independent structural elements simultaneously. It automatically attempts to arrange the RNA fragments in three-dimensional space in a way that yields the best molecular replacement solution (and therefore best phase estimate). To make description of the process more concrete, Fig. 1 depicts an example shell script in which four substructural elements, represented by four independent PDB files generated in COOT (called sl1.pdb, sl2.pdb, sl3.pdb and hx4.pdb), are used simultaneously for automatic molecular replacement (MODE MR_AUTO). A single "native" dataset is read by the program, and all the data between 25.0 and 4.0 Å resolution are employed. (Higher resolution data can be incorporated, if initial attempts fail, but it slows the calculation). Four "ENSEmble" entries are required for the four substructure pdb files, four "COMPosition NUCLeic" entries are required to designate these as nucleic acids, and to assign them molecular weights (based upon their sequences), and four "SEARch ENSEmble" entries are required to designate each as an independent, simultaneous search model.

If all goes well, *PHASER* will output a pdb file and phase set corresponding to the most probable molecular replacement solution. The statistics for this solution will invariably be quite poor compared to what one might expect for standard molecular replacement, and within the present context are essentially meaningless. What is far more important is how the map appears, as this alone gives the most important indication of whether the procedure is beginning to work. Specifically, the *PHASER*-calculated sigma-A-weighted 2Fo-Fc map will show weak or broken up density where the model is incorrect, and more convincing density where the model is approximately correct. Typically, about ½–2/3 of the model will occupy reasonably strong density, and about 1/3 of the model will occupy weak or non-existent density. Again

using *COOT*, the initial model can be edited manually. The most important form of manual intervention at this point of the procedure is to delete mercilessly any nucleotide that does not occupy reasonably strong electron density in a sigma-A-weighted 2Fo–Fc map contoured at 1σ. In addition, portions of any two pdb files that are involved in a steric clash (*i.e.*, that try to occupy the same space in violation of van der Waals repulsion) should be deleted or adjusted to accommodate the electron density. When the editing process is complete, there should be few if any atoms that do not occupy electron density, and no steric clashes should remain. It is however most likely that there is no plausible physical connectivity between subsets of the RNA sequence. This is because the sequences themselves are not enough for the molecular replacement procedure to distinguish between one helix and another. It is in fact not necessary for the starting model to possess the correct sequence; all that is required is that each structural element represents an approximately correct secondary structure. (As an extreme test of this assertion, a solution of a 70 nucleotide RNA crystal structure was obtained using four randomly-generated five-base-paired helices, with no prior knowledge of the actual RNA sequence.)

The edited molecular replacement solution is then refined, typically using *REFMAC* [16] within *COOT* [13,14], and used as a partial model for subsequent iterations of molecular replacement within *PHASER* [15]. At this point simply including one additional helical element is usually sufficient for further model improvement; each addition requires further manual editing as described in the previous paragraph [10]. When further addition of helical elements yields no further improvement in the electron density map, the initial structure is refined using *REFMAC*, and the resulting phase probability distributions are converted to Hendrickson–Lattmann coefficients using the *CCP4* program *HLTOFOM* [17] or the corresponding module within *CNS* [18]. These phases, when combined with the experimentally measured amplitudes, may then be treated as if they were determined by isomorphous replacement, with accompanying phase error estimates. Specifically, improvement of the phases using solvent-flattening (in solvent-flipping mode [19]) within *CNS* [18] will simultaneously reduce model bias and improve the electron density map. The initial model used to generate the phases at this point is discarded.
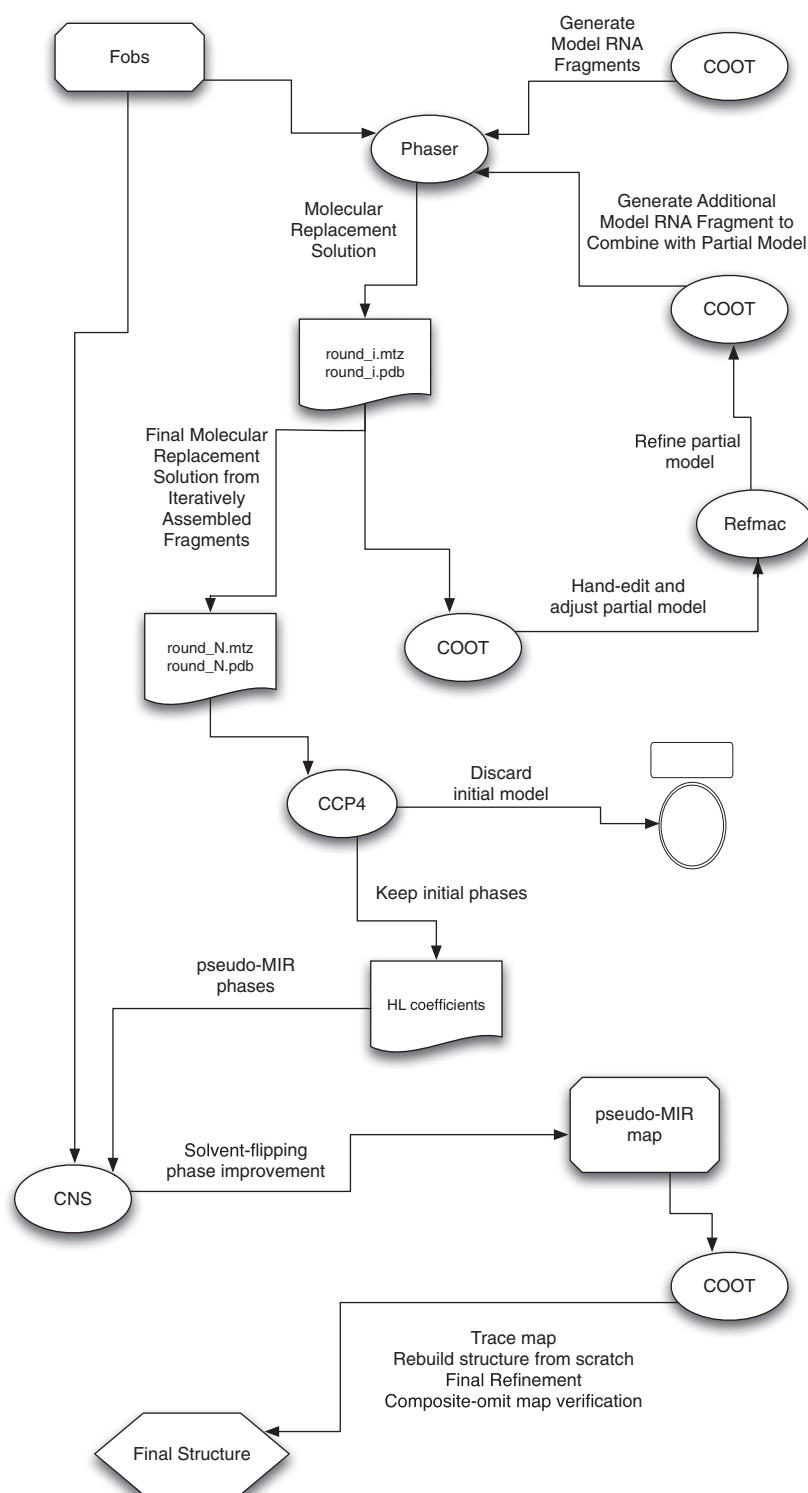
The newly solvent-flattened electron density map may now be treated as if it is an initial experimental map, and a poly-C nucleotide chain can be built into the density using COOT. Simulated-annealing refinement of the initial poly-C structure within *CNS*, using all of the available amplitude data, will likely produce a significantly improved map, one that can be traced using the actual nucleotide sequence. The complete procedure is outlined schematically in Fig. 2, and the initial and final, solvent-flattened, electron density maps are shown in Fig. 3.

2. Software

The procedure outlined in the previous section makes use of a variety of readily available crystallographic software suites. These, along with their specific functionality and the purpose for their use, are described in Table 1.

3. Troubleshooting

As with any molecular replacement procedure, the primary source of complications is model bias. Because the phases are not experimentally determined by model-free approaches, model bias is inherent. We have found the following approaches help to minimize the harmful effect of model bias [2].

```
#!/bin/zsh -ef

# Example shell script to enable simultaneous molecular replacement
# using four independent RNA substructural elements (stem-loops 1,2,3
# and helix 4).

phaser << eof
TITLe sl1 sl2 sl3 automatic
MODE MR_AUTO
HKLIn native_unique.mtz
LABIn F=F_foo SIGF=SIGF_foo
RESOlution 25.0 4.0
ENSEmble sl1 PDB sl1.pdb IDENtity 99
ENSEmble sl2 PDB sl2.pdb IDENtity 99
ENSEmble sl3 PDB sl3.pdb IDENtity 99
ENSEmble hx4 PDB hx4.pdb IDENtity 99
COMPosition NUCLeic MW 11972 NUM 1 #sl1
COMPosition NUCLeic MW 3556  NUM 1 #sl2
COMPosition NUCLeic MW 3521  NUM 1 #sl3
COMPosition NUCLeic MW 3375  NUM 1 #hx4
SEARch ENSEmble sl1
SEARch ENSEmble sl2
SEARch ENSEmble sl3
SEARch ENSEmble hx4
ROOT model_sl1_sl2_sl3_hx4
eof
```

Fig. 1. Using *PHASER* with four independent helical fragments. A shell script to use *PHASER* for the initial molecular replacement calculation is shown.

Fig. 2. Schematic representation of the procedure. An initial assembly of modeled RNA fragments (helices, loops) is used, in combination with a native data set (Fobs), to obtain a starting molecular replacement solution in *PHASER* [15] (round_1.mtz and round_1.pdb). This solution is examined in *COOT* [13,14], and all steric clashes are removed by manual editing, and any nucleotide that does not reside in strong electron density in the sigma-A-weighted 2Fo–Fc map is manually excised. The remaining model is then positionally refined using *REFMAC* [16] (which can be done within *COOT*) and any unoccupied density is modeled and refined in *COOT*. In addition, a new helical fragment (typically 5 base-pairs of A-from helix) for a subsequent iteration of molecular replacement is generated. Using the edited and refined model as a fixed partial model, the next (*i*th) round of molecular replacement in *PHASER* and subsequent editing and refinement is carried out with the new helical fragment. This cycle is repeated until no more helical fragments can be added. At this point, at the *N*th cycle, the best molecular replacement solution (round_N.mtz and round_N.pdb) is used to generate a solvent-flattened electron density map in *CNS* [18]. The model (round_N.pdb) is discarded, and the calculated phase probability distributions are converted to Hendrickson–Lattmann coefficients within *CCP4* [17] and are then imported into *CNS*, along with the native data set (Fobs), and treated as if they were experimentally determined MIR phases to be solvent-flattened. The resulting map is then used for building the final model from scratch (*i.e.*, without reference to the discarded molecular replacement solution round_N.pdb) as if the map were derived from experimental MIR phases. The resulting structure is then checked against a composite–omit map generated in *CNS* in which 10% of the model is omitted from each element of the composite, and phases are re-generated from a standard simulated annealing procedure with a starting temperature of 4000 K.
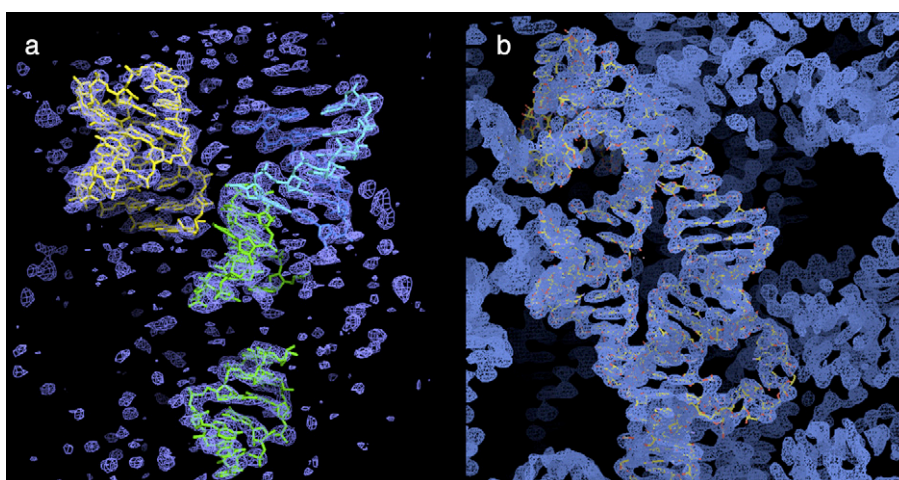
Fig. 3. Electron density maps. (a) Initial molecular replacement solution with four secondary structural fragments of the L1 ligase ribozyme. (b) The solvent-flattened electron density "pseudo-MIR" map created within *CNS*. A final refined model of the L1 ligase (2oiu) is superimposed on one asymmetric unit of electron density.

Table 1
Software of use for solving RNA structures by piecewise molecular replacement.

Suggested software	Functionality	Purpose
ViennaRNA	RNAfold	Predict secondary structure of RNA
COOT	Modelling tools > ideal DNA/RNA	Create ideal A-form RNA helical fragments for initial molecular replacement
	Molecular editing and structural manipulation	Delete or adjust portions of initial molecular replacement solution within weak or non-existent density Rebuilding of final model into pseudo-MIR map
PHASER	Automated molecular replacement	Position the various helical fragments correctly in 3D space to produce an approximate initial phase set
REFMAC	Conventional crystallographic refinement	Optimize molecular geometry subsequent to editing
CCP4	Data manipulation	Create a pseudo-experimental phase set and convert to Hendrickson–Lattman coefficients for solvent-flattening and partial model refinement
	Generation of Hendrickson–Lattmann coefficients	
CNS or *PHENIX*	Solvent-flattening	Improve initial phase estimate and reduce model bias using solvent flipping/flattening
	Blurring Hendrickson–Lattmann coefficients	Further reduce model bias via "blurring" HL coefficients
	Simulated annealing crystallographic refinement	Calculation of "pseudo-MIR" electron density map
	Composite–Omit map calculation	Simulated-annealing refinement of final model Composite–omit map calculation to check Structural veracity

3.1. Blurring of Hendrickson–Lattmann coefficients

Solvent-flattening and other density-modification procedures artificially inflate the figure of merit for estimated phases. Hence many partial models that have been crystallographically refined using modern approaches that attempt to model solvent effects will have phase probability estimates that are unrealistically high. This becomes problematic if the estimated phases are precise but rather inaccurate, as will be the case with the piecewise molecular replacement procedure we have described. The initial phases may be weighted too strongly and will thus not benefit from further attempts at improvement. The program *CNS* [18] is distributed with a module that permits the HL coefficients to be "blurred" by manually attenuating the temperature and scale factors (hlco-eff_blur.inp). Doing so inhibits a partial model from "taking over" density-modification procedures as a result of model bias, and can thus greatly improve the quality of the pseudo-experimental electron density map [20].

3.2. Phase perturbation in EDEN [21]

The real-space density modification program EDEN uses a holographic procedure to minimize model bias, and further permits the user to randomly perturb initial phase estimates to test for reconvergence. An EDEN map generated from a partial model is often robust enough to enable construction of 2/3 of a missing asymmetric unit.

3.3. Composite–Omit maps in CNS

Systematic elimination of 10% of a structure followed by simulated-annealing refinement of the remaining structure permits reconstruction of an electron density map corresponding to the omitted 10% of the structure. Ten such unique maps, when composited, will produce a composite–omit map that has minimal model bias. This procedure is easily implemented within *CNS* [18] and should always be employed as a reality check.

4. Concluding remarks

Contrary to initial assumptions, it is possible to use molecular replacement to solve crystal structures of RNAs having unique folds without prior knowledge of their tertiary structures. This approach appears to be possible due to the rather high degree of regularity of RNA secondary structures (such as A-form helices and tetraloops) and the comparative ease with which they may be pre-

dicted to form based upon known sequences. This in turn implies that, even in the absence of crystallographic phase information, there is sufficient information in a single crystal diffraction pattern to deduce the arrangement of secondary structural elements in three-dimensional space. In practice, this can be accomplished using simultaneous molecular replacement of several RNA helical fragments.

References

[1] G. Taylor, Acta Crystallogr. D Biol. Crystallogr. 59 (2003) 1881–1890.
[2] R.J. Read, Acta Crystallogr. D Biol. Crystallogr. 57 (2001) 1373–1382.
[3] Y.W. Chen, E.J. Dodson, G.J. Kleywegt, Structure 8 (2000) R213–R220.
[4] B. Qian, S. Raman, R. Das, P. Bradley, A.J. McCoy, R.J. Read, D. Baker, Nature 450 (2007) 259–264.
[5] A.Y. Keel, R.P. Rambo, R.T. Batey, J.S. Kieft, Structure 15 (2007) 761–772.
[6] M.P. Robertson, W.G. Scott, Science 315 (2007) 1549–1553.
[7] Y.I. Chi, M. Martick, M. Lares, R. Kim, W.G. Scott, S.H. Kim, PLoS Biol. 6 (2008) e234.
[8] D.J. Klein, T.E. Edwards, A.R. Ferré-D'Amaré, Nat. Struct. Mol. Biol. 16 (2009) 343–344.
[9] N. Kulshina, N.J. Baird, A.R. Ferré-D'Amaré, Nat. Struct. Mol. Biol. 16 (2009) 1212–1217.
[10] M.P. Robertson, W.G. Scott, Acta Crystallogr. D Biol. Crystallogr. 64 (2008) 738–744.
[11] M. Zuker, Nucleic Acids Res. 31 (2003) 3406–3415.
[12] I.L. Hofacker, M. Fekete, P.F. Stadler, J. Mol. Biol. 319 (2002) 1059–1066.
[13] P. Emsley, K. Cowtan, Acta Crystallogr. D Biol. Crystallogr. 60 (2004) 2126–2132.
[14] P. Emsley, B. Lohkamp, W.G. Scott, K. Cowtan, Acta Crystallogr. D Biol. Crystallogr. 66 (2010) 486–501.
[15] A.J. McCoy, R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, R.J. Read, logo 40 (2007) 658–674.
[16] G.N. Murshudov, A.A. Vagin, E.J. Dodson, Refinement of Macromolecular Structures by the Maximum-Likelihood Method 53 (1997) 240–255.
[17] M.D. Winn, J. Synchrotron. Radiat. 10 (2003) 23–25.
[18] A.T. Brunger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren, Acta Crystallogr. D Biol. Crystallogr. 54 (Pt 5) (1998) 905–921.
[19] J.P. Abrahams, A.G. Leslie, Acta Crystallogr. D Biol. Crystallogr. 52 (1996) 30–42.
[20] T.A. Jones, J.Y. Zou, S.W. Cowan, Acta Crystallogr. A 47 (Pt 2) (1991) 110–119.
[21] A. Szoke, Acta Crystallogr. A 57 (2001) 586–603.