BIOCHEMISTRY
including biophysical chemistry & molecular biology
Rapid Report

DOI: 10.1021/bi100142y

Robust and Generic RNA Modeling Using Inferred Constraints: A Structure for the Hepatitis C Virus IRES Pseudoknot Domain[†]

Christopher A. Lavender,[‡] Feng Ding,[§] Nikolay V. Dokholyan,*[,§] and Kevin M. Weeks*[,‡]

[‡]*Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, and* [§]*Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina 27599-7260*

Received January 28, 2010; Revised Manuscript Received May 12, 2010

ABSTRACT: RNA function is dependent on its structure, yet three-dimensional folds for most biologically important RNAs are unknown. We develop a generic discrete molecular dynamics-based modeling system that uses long-range constraints inferred from diverse biochemical or bioinformatic analyses to create statistically significant ($p < 0.01$) nativelike folds for RNAs of known structure ranging from 45 to 158 nucleotides. We then predict the unknown structure of the hepatitis C virus internal ribosome entry site (IRES) pseudoknot domain. The resulting RNA model rationalizes independent solvent accessibility and cryo-electron microscopy structure information. The pseudoknot domain positions the AUG start codon near the mRNA channel and is tRNA-like, suggesting the IRES employs molecular mimicry as a functional strategy.

Critical RNA structures directly regulate gene expression, splicing, and translation (*1*), but the structures of most biologically important RNA folds are currently unknown. Recent studies highlight significant successes in *ab initio* structure prediction of local helical structure and of small RNA motifs (*2*). However, the ability of current approaches to predict RNA structure accurately decreases rapidly with increasing RNA size. *De novo* prediction of large RNA structures with complex, nontrivial, three-dimensional folds from sequence alone remains beyond the realm of current automated algorithms. A compelling alternative is to develop modeling methods for facile incorporation of readily obtained experimental information.

Long-range constraints for RNA modeling can be inferred from a variety of biochemical and bioinformatic techniques, ranging from chemical footprinting and cross-linking to sequence covariation (*3*). Algorithms devised thus far are making significant progress toward the goal of incorporating specific classes of tertiary structure information into RNA structure refinement (*4*). However, current refinement approaches still make large assumptions about the nature of the constraint information used and are closely tied to the specific techniques employed to infer long-range interactions.

To address these challenges, we develop a generic and efficient approach for accurately predicting RNA folds using tertiary structure information as inferred from diverse biochemical or bioinformatic techniques. Distance constraints are incorporated into a discrete molecular dynamics (DMD) engine (*2b*) that uses a single refinement approach for all classes of tertiary structure constraint information. RNA nucleotides are represented as three pseudoatoms corresponding to the phosphate (P), sugar (S), and base (B) moieties (Figure 1A). Three pseudoatoms are sufficient for the development of nucleotide-resolution RNA models with rigid base-paired helices and physically meaningful base stacking interactions, while still allowing large RNAs to be refined efficiently.

Inferred pairwise tertiary constraints are incorporated via a generic constraint system that uses a potential well with an effective length of 15.0 Å and a depth of 2.0 kcal/mol between base pseudoatoms (Figure 1B). This constraint system is compatible with techniques that do not directly provide distance information but instead merely imply pairwise interactions, as with mutational studies.

Four RNAs were selected to benchmark constrained structure refinement: domain III of the cricket paralysis virus internal ribosome entry site (CrPV) (49 nucleotides), a full-length hammerhead ribozyme from *Schistosoma mansoni* (HHR) (67 nucleotides), *Saccharomyces cerevisiae* $\mathrm{tRNA^{Asp}}$ (75 nucleotides), and the P546 domain of the *Thermus thermophilia* group I intron (P546) (158 nucleotides). Each of these RNAs has a complex three-dimensional fold dependent both on local helical structure and on long-range tertiary interactions. Prior to publication of the high-resolution structures (*5*), significant biochemical or bioinformatic data describing tertiary interactions were available for each RNA. The secondary structure was also known with good accuracy in each case. Only this prior information (Table S1 of the Supporting Information) was used during refinement.

A single generic and completely automated refinement protocol was applied to each RNA. Simulations begin with the RNA strand in an extended conformation at a high temperature. Constraints based on the secondary structure are included, and the molecular system is annealed to allow helices to form. Constraints for inferred tertiary interactions are incorporated, and the RNA is cooled to a final target temperature. RNA structures from this step (100000) are subjected to automated clustering. The centroid of the most populated cluster is selected as the final predicted structure. Given our refinement model, this structure is representative of the lowest-free energy state.

Refined models for all four test RNAs are accurate (Figure 2). The root-mean-square deviations (RMSDs) of the phosphate backbone relative to the accepted structures for the CrPV, HHR, $\mathrm{tRNA^{Asp}}$, and P546 RNAs were 3.6, 5.4, 6.4, and 11.3 Å, respectively. Analysis of the RNA structure prediction significance (p value) (*6*) shows that the probabilities that these models result from chance are small (2×10^{-3}, 2×10^{-5}, 3×10^{-6}, and $\leq 10^{-6}$, respectively).

[†]This work was supported by National Institutes of Health Grant GM064803 (to K.M.W.) and an ARRA supplement (to K.M.W. and N.V.D.).

*To whom correspondence should be addressed. E-mail: weeks@unc.edu (K.M.W.) and dokh@med.unc.edu (N.V.D.). Phone: (919) 962-7486. Fax: (919) 962-2388 (K.M.W.).

Published on Web 05/27/2010

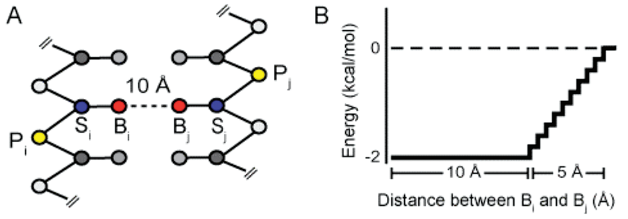

FIGURE 1: Generic constraint system. (A) Three-bead model for RNA. (B) Interaction potential for distance constraints.

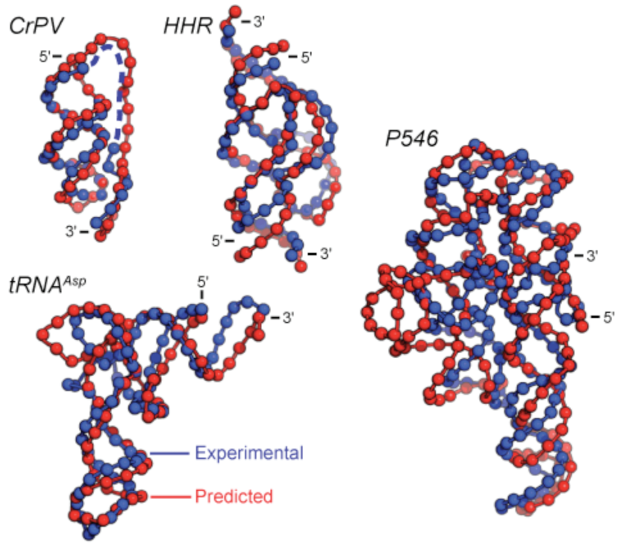

FIGURE 2: Comparison of predicted and experimental RNA structures. Spheres indicate phosphate groups. The root-mean-square deviations for the CrPV, HHR, tRNA, and P546 RNAs are 3.6, 5.4, 6.4, and 11.3 Å, respectively.

There are two critical results from this analysis of RNAs with known structures. First, nativelike RNA folds were obtained in every case despite the diversity of structural information used to constrain refinement (Table S1 of the Supporting Information). Second, prediction quality was maintained as RNA size increased from a 49-nucleotide pseudoknot to a 158-nucleotide RNA domain with a complex tertiary structure (Figure 2).

Our approach compares favorably to other coarse-grained RNA modeling approaches. Folds for tRNA^{Phe} and the P546 domain have been predicted with the program NAST in which each RNA nucleotide is represented by a single pseudoatom (*4d*). NAST modeling was constrained using structure information similar to that used in our refinements. Of the resulting models, the most accurate had RMSDs relative to the accepted structure of 8.0 and 16.3 Å for tRNA and P546, respectively, whereas our approach yields smaller RMSDs of 6.4 and 11.3 Å, respectively. NAST simulations used 300 h per RNA, as compared to 18-40 real-time computing hours for the DMD-based refinements (Figure 2). These comparisons highlight both the accuracy and efficiency of our constrained DMD approach.

Having shown that this fully automated approach recapitulates nativelike folds for diverse, well-characterized RNAs, we sought to apply this algorithm to an RNA for which extensive biochemical information exists but whose structure is unknown. We focused on the pseudoknot domain in the hepatitis C virus (HCV) internal ribosome entry site (IRES).

IRES elements bypass canonical cap-dependent eukaryotic translation initiation by directly recruiting ribosomes to internal sequences in a mRNA (*7*). Structural studies have significantly improved our understanding of functional mechanisms of IRES elements (*5d*, *8*). High-resolution structures are available for many elements of the HCV IRES (*9*); however, the three-dimensional fold for the pseudoknot domain (HCV-PK) has not been determined. The pseudoknot domain consists of a pseudoknot at the base of domain III (dIII) and its flanking structures (Figure 3A). Mutation of the pseudoknot inhibits translation initiation in HCV replication (*10*). Compensatory mutations that restore the pseudoknot do not always restore HCV translation activity, suggesting that sequence conservation is required for functions beyond base pairing. The pseudoknot domain contains the AUG start codon for translation of the HCV polypeptide (yellow in Figure 3A). Solvent accessibility experiments show the pseudoknot domain is the most highly structured element in the IRES (*11*). Extensive available biochemical information and intense biomedical interest make the HCV-PK RNA an ideal candidate for deriving biological insights based on structural modeling.

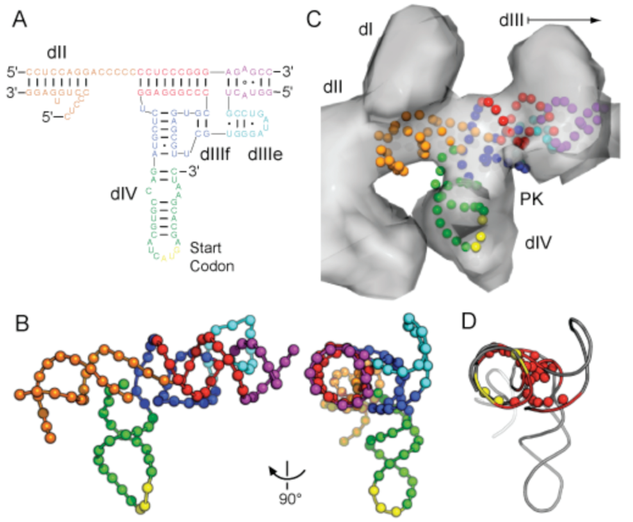

FIGURE 3: RNA structure prediction for the pseudoknot domain of the HCV IRES. (A) Secondary structure (*10*). (B) Predicted structure of the uncomplexed pseudoknot domain. Spheres indicate phosphate positions; nucleotides are colored as shown in panel A. (C) HCV-PK model placed into the electron density of the IRES-ribosome complex (*12b*). (D) Superposition of hydroxyl radical protection data (*11*) on the HCV-PK RNA model. Spheres correspond to protected sugar pseudoatoms; red and yellow indicate strong and moderate protection, respectively.

A three-dimensional model for the HCV-PK domain RNA was refined using the same fully automated folding algorithm as for the four test RNAs. Base pairs in the pseudoknot were modeled as generic tertiary constraints.

The predicted HCV-PK structure is dominated by two structural features (Figure 3B). The first is the four-way junction comprised of stems at the base of dIII (red and purple), dIIIe (cyan), and dIIIf (blue). The second consists of base stacking interactions between the pseudoknot (blue) and dIV (green). The nucleotide linkages between these two motifs are short and lock the dIV helix in a conformation perpendicular to the plane described by the helices of the four-way junction.

Two classes of independent experiments support the proposed structure for the HCV-PK RNA. First, the predicted tight RNA folding in the four-way junction and pseudoknot is supported by protection from hydroxyl radical cleavage, indicative of solvent inaccessible regions of the RNA backbone (*11*). Solvent

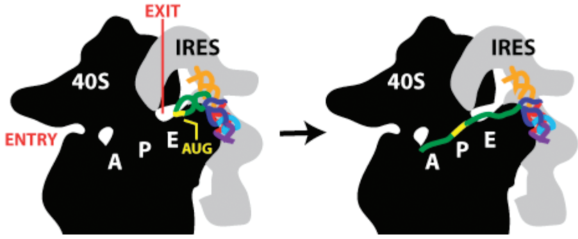

FIGURE 4: Docking of HCV IRES RNA into the mRNA channel of the 40S ribosome. Cartoons of the 40S subunit (black) and HCV IRES (gray) are based on cryo-EM studies (*12*). The AUG start site codon, mRNA entry and exit sites, and tRNA binding sites are labeled on the 40S subunit. The HCV-PK model is colored and positioned in the same orientation, relative to the cryo-EM density, as in Figure 3C.

inaccessible regions fall precisely in the interior of the four-way junction and at the interface of this element with the pseudoknot (red and yellow spheres in Figure 3D).

Second, the HCV-PK model is consistent with cryo-EM electron density maps of the IRES−ribosome complex. Our model of the HCV-PK is that of the uncomplexed IRES, and conformational changes occur in both the ribosome and IRES when the IRES interacts with ribosomal subunits and translation initiation factors (*12*). For example, domain IV likely unfolds to allow positioning of the start codon in the P-site (*13*) (see the Supporting Information). Nevertheless, the core of our model fits well in the density assigned to the pseudoknot domain in the cryo-EM electron density maps of the IRES−ribosome complex (*12b*). The critical correlations are that dII and dIII (orange and purple, respectively, in Figure 3C and Figure S1 of the Supporting Information) are positioned to connect sensibly with the rest of the IRES, and the perpendicular orientation of the pseudoknot (blue) allows the AUG start codon in dIV (yellow in Figures 3C and 4) to be positioned in or near the mRNA channel. Our HCV-PK model also fits well with high-resolution IRES structures positioned in the cryo-EM density (Figure S2 of the Supporting Information).

Several functional hypotheses are consistent with the predicted model. First, the HCV-PK RNA is L-shaped, similar to tRNA, and can be aligned with yeast tRNA^{Asp} (not shown). Formation of a tRNA-like structure is consistent with biochemical studies showing that the HCV IRES is cleaved by the tRNA-recognizing ribonuclease RNase P (*14*). tRNA mimicry also rationalizes the presence of a seven-nucleotide loop at the end of domain IV, a structural feature that is generally uncommon in RNA but present in the anticodon loops of most tRNAs.

A recent structural study also yielded evidence of tRNA mimicry in domain III of the CrPV IRES (*5d*). Though the HCV-PK model and CrPV experimental structure have distinct folds, both support tRNA mimicry as a common strategy employed by IRES structures and are consistent with extensive examples of tRNA mimicry in biologically diverse RNAs (*15*).

Second, the perpendicular orientation of the pseudoknot relative to the four-way junction may function to position the AUG start codon for translation initiation. In cryo-EM maps of both the 40S− and 80S−IRES complexes, density corresponding to the pseudoknot domain is adjacent to the channel occupied by the mRNA template during translation (*12*). Thus, dIV and, specifically, the AUG start codon will be positioned near the ribosome mRNA exit site.

These observations support a model in which the IRES pseudoknot domain docks initially near the ribosome exit channel, facilitated by its tRNA-like structure (Figure 4, left). Our model suggests additional conformational changes are required in the IRES and ribosome for the AUG start codon to fully occupy the mRNA channel. A modest unfolding of the dIV helix would then allow this element to serve as the mRNA template for translation of the HCV polyprotein (Figure 4, right).

RNA structure refinement using inferred constraints consistently yields nativelike models for RNAs spanning 49−158 nucleotides. This approach does not require a specific optimized form for the long-range constraints but does require knowledge of through-space tertiary interactions. The success of this approach implies that knowledge of only a few long-range constraints is sufficient to refine accurate folds for many RNAs with complex structures.

The HCV-PK domain model rationalizes substantial preexisting biochemical information for this RNA and provides specific and novel functional insights useful for guiding future hypotheses and experiments. RNA structure refinement using inferred constraints holds significant promise for understanding the functions of many biologically important RNAs whose analysis is recalcitrant to high-resolution approaches.

SUPPORTING INFORMATION AVAILABLE

Details of the refinement algorithm, two figures, biochemical data used to constrain refinements, discussion of placing the HCV-PK model in the cryo-EM density, and PDB files of predicted RNA models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Gesteland, R. F., et al. (2006) The RNA World, 3rd ed., Cold Spring Harbor Laboratory Press, Plainview, NY.
2. (a) Das, R., and Baker, D. (2007) *Proc. Natl. Acad. Sci. U.S.A. 104*, 14664–14669. (b) Ding, F.; et al. (2008) *RNA 14*, 1164–1173. (c) Parisien, M., and Major, F. (2008) *Nature 452*, 51–55.
3. (a) Ziehler, W. A., and Engelke, D. R. (2001) Current Protocols in Nucleic Acid Chemistry, Chapter 6, Unit 6, p 1, Wiley-Interscience, New York. (b) Juzumiene, D.; et al. (2001) *Methods 25*, 333–343. (c) Gutell, R. R.; et al. (1992) *Nucleic Acids Res. 20*, 5785–5795.
4. (a) Badorrek, C. S.; et al. (2006) *Proc. Natl. Acad. Sci. U.S.A. 103*, 13640–13645. (b) Das, R.; et al. (2008) *Proc. Natl. Acad. Sci. U.S.A. 105*, 4144–4149. (c) Gherghe, C. M.; et al. (2009) *J. Am. Chem. Soc. 131*, 2541–2546. (d) Jonikas, M. A.; et al. (2009) *RNA 15*, 189–199.
5. (a) Westhof, E.; et al. (1988) *Acta Crystallogr. A44*, 112–123. (b) Cate, J. H.; et al. (1996) *Science 273*, 1678–1685. (c) Martick, M., and Scott, W. G. (2006) *Cell 126*, 309–320. (d) Costantino, D. A.; et al. (2008) *Nat. Struct. Mol. Biol. 15*, 57–64.
6. Hajdin, C. E., Ding, F., Dokholyan, N. V., and Weeks, K. M. (2010) *RNA 16*, (in press).
7. (a) Pisarev, A. V.; et al. (2005) *C. R. Biol. 328*, 589–605. (b) Kieft, J. S. (2008) *Trends Biochem. Sci. 33*, 274–283.
8. Pfingsten, J. S.; et al. (2006) *Science 314*, 1450–1454.
9. (a) Lukavsky, P. J. (2008) *Virus Res. 139* (2), 166–171. (b) Filbin, M. E., and Kieft, J. S. (2009) *Curr Opin Struct Biol 19*, 267–276.
10. (a) Wang, C.; et al. (1995) *RNA 1*, 526–537. (b) Kieft, J. S.; et al. (2001) *RNA 7*, 194–206.
11. Kieft, J. S.; et al. (1999) *J. Mol. Biol. 292*, 513–529.
12. (a) Spahn, C. M.; et al. (2001) *Science 291*, 1959–1962. (b) Boehringer, D.; et al. (2005) *Structure 13*, 1695–1706.
13. (a) Pestova, T. V.; et al. (1998) *Genes Dev. 12*, 67–83. (b) Otto, G. A., and Puglisi, J. D. (2004) *Cell 119*, 369–380. (c) Fraser, C. S.; et al. (2009) *Nat. Struct. Mol. Biol. 16*, 397–404.
14. (a) Nadal, A.; et al. (2002) *J. Biol. Chem. 277*, 30606–30613. (b) Lyons, A. J., and Robertson, H. D. (2003) *J. Biol. Chem. 278*, 26844–26850.
15. Hammond, J. A.; et al. (2009) *RNA 15*, 294–307.