Supporting Information for:

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*

*correspondence, weeks@unc.edu and dokh@med.unc.edu

RNA Sequences.

CrPV: 5'-UGCGG UUUUU CAGAU UAGGU AGUCG AAAAA CCUAA GAAAU UUACC UGCU-3'

HHR: 5'-GGAUG UACUA CCAGC UGAUG AGUCC CAAAU AGGAC GAAAC GCCAA AAGGC GUCCU GGUAU CCAAU CC-3'

$tRNA^{Asp}$: 5'-UCCGU GAUAG UUUAA UGGUC AGAAU GGGCG CUUGU CGCGU GCCAG AUCGG GGUUC AAUUC CCCGU CGCGG AGCCA-3'

P546: 5'-GAAUU GCGGG AAAGG GGUCA ACAGC CGUUC AGUAC CAAGU CUCAG GGGAA ACUUU GAGAU GGCCU UGCAA AGGGU AUGGU AAUAA GCUGA CGGAC AUGGU CCUAA CCACG CAGCC AAGUC CUAAG UCAAC AGAUC UUCUG UUGAU AUGGA UGCAG UUC-3'

HCV-PK: 5'-CUCCC CUGUG AGGUU UUCCU CCAGG ACCCC CCCUC CCGGG AGAGC CUUUU GGUAC UGCCU GAUAG GGUGC UUGCG AGUGC CCCGG GAGGU CUCGU AGACC GUGCA UCAUG AGCAC GAAUC-3'

Sequences correspond to those used in prior structural studies (*1-4*). In HHR, where the crystallographic model includes two strands, the strands were connected with a 5'-AAAA-3' linker. The HCV-PK sequence was taken from genotype 1b (*5*). The HCV-PK model includes nucleotides 40-52, 111-139, and 285-354, comprising the base of domain II, the four-way junction at the base of domain III, and domain IV. During modeling, these three segments of the HCV IRES sequence were connected using two 5'-UUUU-3' linkers.

Secondary and tertiary structure reference information. Sources for the secondary structures and tertiary contact information used to constrain DMD refinement are outlined in detail in Table S1. Base pairs in the secondary structures were constrained as described previously (*6, 7*). Tertiary contacts were imposed using the generic constraint system created in this work.

In the case of the CrPV RNA, tertiary constraints were used to describe the pseudoknot identified using chemical modification techniques and mutational analysis (*1, 8*).

For $tRNA^{Asp}$, pair-wise tertiary contacts were inferred from covariation studies summarized by Gutell and colleagues (*9*). These contacts are consistent with the tertiary interactions originally compiled by Levitt (*3*).

In the case of the HHR RNA, the specific nucleotides involved in the loop-loop interaction in the full-length hammerhead ribozyme were not known prior to crystallization. To approximate this interaction given the incomplete level of pre-existing information, generic constraints were included between residues in the middle of each loop in the regions generally known to interact (*10, 11*). To avoid biasing

the orientation of the loops in this long-range interaction, constraints were included for all pair-wise combinations for each side of each loop, even though one-half of these contacts were likely to be (partially) incorrect. The resulting quality of the prediction for the HHR RNA emphasizes that the approach taken in this work is tolerant of modest misassignments of the specific partners participating in long-range tertiary interactions.

For the P546 domain, tertiary constraints were based on two base triples inferred from mutational analysis and DMS modification (*12*) and from mutational analysis and hydroxyl radical protection data (*13*).

In the case of HCV-PK, tertiary constraints were based upon the pseudoknot inferred from mutational analysis, chemical and enzymatic probing, and thermodynamic calculations (*14*).

DMD algorithm. The DMD engine (*6, 7*) models each nucleotide as three pseudo-atoms corresponding to the phosphate, sugar, and base moieties. Pair-wise interactions including base pairing, base stacking, packing interactions, and electrostatic repulsion are approximated using square-well potentials (*7*). In base-paired regions of the model, distance constraints between base and phosphate pseudo-atoms are used to maintain the rigid structure characteristic of the RNA double helix (*6*).

The engine used in this work extends base stacking interactions to both base-paired and single-stranded RNA regions. Single-stranded stacking interactions contribute $-k_BT$ to the RNA free energy, where k_B is the Boltzmann constant and T equals 300 K ($-k_BT$ = -0.6 kcal/mol). This is one-half of the free energy contribution assigned to base-paired stacking interactions ($-2k_BT$) (*6*).

Generic constraint system. Distance constraints were included between the base pseudo-atoms of residues inferred to participate in pair-wise interactions. A maximum free energy bonus of 2.0 kcal/mol was applied when base pseudo-atoms were within 10.0 Å of each other. For each 0.5 Å beyond the inter-pseudo-atom distance of 10.0 Å, the bonus was reduced by 0.2 kcal/mol, giving the constraint an effective length of 15.0 Å. The attractive potential is described in Figure 1 and in the following potential function, where d is the distance between base pseudo-atoms:

$$E_{\text{constraint}} = \begin{cases} -2.0\,\text{kcal/mol}, 0 \le d < 10.0\,\text{Å} \\ -1.8\,\text{kcal/mol}, 10.0 \le d < 10.5\,\text{Å} \\ \vdots \\ -0.2\,\text{kcal/mol}, 14.5 \le d < 15.0\,\text{Å} \\ 0, 15.0\,\text{Å} \le d \end{cases}$$

General RNA refinement protocol. Simulations begin with the RNA strand in an extended linear conformation at a high temperature. The RNA sequence and constraints based on secondary structure were the initial input to the DMD algorithm. The RNA was first subjected to a folding phase designed to allow base pairs and local helical structure to form. In this phase, the RNA was cooled through the following automated steps, where T_i and T_f are the initial and final reduced temperatures [in kcal/(mol × k_B)]: (1) T_i, T_f = 30, for 10^5 DMD time units (tu); (2) T_i = 0.30, T_f = 0.28, 2 × 10^4 tu; (3) T_i, T_f =0.28, 10^5 tu; (4) T_i =0. 28, T_f = 0.26, 2 × 10^4 tu; (5) T_i, T_f = 0.26, 10^5 tu; (6) T_i = 0.26, T_f = 0.24, 2 × 10^4 tu; (7) T_i, T_f = 0.24, 10^5 tu; (8) T_i = 0.24, T_f = 0.22, 2 × 10^4 tu; (9) T_i, T_f = 0.22, 10^5 tu; (10) T_i, T_f = 0.22, 2 × 10^4 tu. After step 10, constraints describing tertiary contacts were added. The RNA model was then cooled to a target reduced temperature through the following steps: (11) T_i, T_f = 0.22, 10^5 tu; (12) T_i = 0.22, T_f = 0.20, 2 × 10^4 tu; (13) T_i, T_f = 0.20, 10^5 tu; (14) T_i = 0.20, T_f = 0.18, 2 × 10^4 tu; (15) T_i, T_f = 0.18, 10^5 tu; (16) T_i = 0.18, T_f = 0.16, 2 × 10^4 tu; (17) T_i, T_f = 0.16, 10^5 tu; (18) T_i = 0.16, T_f = 0.14, 2 × 10^4 tu; (19) T_i, T_f = 0.14, 10^5 tu; (20) T_i = 0.14, T_f = 0.12, 2 × 10^4 tu; (21) T_i, T_f = 0.12, 10^5 tu; (22) T_i = 0.12, T_f = 0.10, 2 × 10^4 tu; (23) T_i, T_f = 0.10, 10^5 tu. 100,000 structures are generated at this final refinement step. To select a

representative structure for each refinement, structures from step 23 were subjected to hierarchical clustering (*15*), as described (*6*). Structures were first filtered on the basis of energy and simulation distance. Clustered structures were required to have an energy less than the median energy from step 23 and were required to be at least 1000 tu apart from other clustered structures (to prevent analysis of consecutive structures). Structures were binned by RMSD value into five clusters. The centroid of the cluster with the highest population was taken to be the representative structure.

Refinements were performed on a Linux workstation (Intel Pentium 4 processor, 3.2 GHz) running Fedora Core 4. Refinement times ranged from 18 (CrPV, 49 nts) to 42 hrs (P546, 158 nts).

In some cases, the local structure that forms before incorporation of tertiary contact constraints restricts conformational space such that residues implicated in a tertiary interaction may not come into contact during refinement. In cases where imposed tertiary contacts were not present in the final structure (HHR, P546 and HCV-PK RNAs), the refinement was repeated with *temporary* strong constraints during step 10; these constraints simply function to promote initial collapse of the RNA molecule. For a given long-range pair-wise interaction, square well potentials were included between each pair of phosphate, sugar, and base pseudo-atoms, providing a graduated potential well that extends from an inter-residue distance of 10.0 to 155.0 Å. The maximum energy bonus for this constraint set was 20.0 kcal/mol when phosphate pseudo-atoms are within 85 Å of each other, 15.0 kcal/mol when sugar atoms are within 30 Å, and 10.0 kcal/mol when bases are within 10 Å.

We emphasize that, subject to the requirement to force compaction if tertiary interaction constraints are not fulfilled, this refinement algorithm is fully automated and does not make use of *ad hoc* user input. The algorithm initially yields 100,000 structures and the representative structure is selected by clustering without knowledge of the accepted crystallographic model. The approach is strongly dependent on the nature of the tertiary structure constraint information available. The four benchmark RNAs were chosen based on the availability of typical classes of biochemical, mutational, and sequence covariation information (Table 1). We anticipate that this approach will work well for other RNAs for which there is comparable information.

RMSD calculations were performed using lsqman (*16*) and include all phosphate pseudo-atom positions.

RNA model images were composed using Pymol (*24*) with the exception of Figure 3D, which was generated using UCSF Chimera (*18*).

Fitting the HCV-PK model and other RNA elements into the cryo-EM electron density map. HCV IRES electron density maps were generously provided by Holger Stark and Daniel Boehringer. Models for the HCV-PK domain alone (Figure 3) or with previously determined high-resolution structures (Figure S1) were placed into the cryo-EM electron density of the IRES in complex with the 80S ribosome (*17*) by manual fitting using UCSF Chimera (*18*). Source files were: domain II (PDB ID 1p5p, orange in Figure S1) (*19*), IIIb (1kp7, yellow) (*20*), and IIIe (1f85, cyan) (*21*), four-way junction IIIabc (1kh6, green) (*22*). Placement of high-resolution structures is consistent with prior models (PDB ID 2agn) (*17*). The fit of the four-way junction IIIabc in the cryo-EM map is ambiguous as has been described previously (*17*).

We have included a direct comparison between our model and the electron density of the IRES in complex with the 80S ribosome (*17*) (Figure 3). We note that uncertainties associated with the cryo-EM map also allow other interpretations; in particular, the conformation of domain IV, as visualized in the ribosome complex by cryo-EM, is not fully understood. The HCV IRES undergoes conformational changes during translation initiation. Domain IV is likely to be unwound at some stage during its association with the ribosome to allow placement of the AUG start codon in the ribosome decoding

channel; however, there is no direct evidence for unwinding in the complexes analyzed by cryo-EM. tRNA is not visible in the cryo-EM map, indicating that the P-site tRNA has dissociated from the original IRES-ribosome complex. Dissociation of the P-site tRNA may cause or be caused by conformational changes in the IRES. The cryo-EM maps are thus consistent either with a base paired stem-loop conformation for domain IV or with a single stranded, extended structure for this domain with additional density in this region corresponding to ribosome components. (Interactions with the IRES may also result in conformational changes in the ribosome which would make attribution of the electron density difficult at the IRES-ribosome interface.) These two options are illustrated in Figure S1.

Independent of the conformation of domain IV, the cryo-EM density is strongly supportive of the HCV-PK model, generated by DMD (Figure 3). The IRES pseudoknot and 4-way junction structural elements fit well and are positioned to connect sensibly with the rest of the IRES (emphasized by colored arrows, Figure S1B). Both domain IV conformations are consistent with the key structural feature of our model which projects domain IV and the AUG start codon from the IRES in a perpendicular orientation that allows positioning into the ribosome decoding site (Figures 4 & S1).

References

1. Jan, E., and Sarnow, P. (2002) Factorless ribosome assembly on the internal ribosome entry site of cricket paralysis virus, *J Mol Biol 324*, 889-902.

2. Canny, M. D., Jucker, F. M., Kellogg, E., Khvorova, A., Jayasena, S. D., and Pardi, A. (2004) Fast cleavage kinetics of a natural hammerhead ribozyme, *J Am Chem Soc 126*, 10848-10849.

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

4. Cech, T. R., Damberger, S. H., and Gutell, R. R. (1994) Representation of the secondary and tertiary structure of group I introns, *Nat Struct Biol 1*, 273-280.

5. Brown, E. A., Zhang, H., Ping, L. H., and Lemon, S. M. (1992) Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs, *Nucleic Acids Res 20*, 5041-5045.

6. Gherghe, C. M., Leonard, C. W., Ding, F., Dokholyan, N. V., and Weeks, K. M. (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics, *J Am Chem Soc 131*, 2541-2546.

7. Ding, F., Sharma, S., Chalasani, P., Demidov, V. V., Broude, N. E., and Dokholyan, N. V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms, *RNA 14*, 1164-1173.

8. Kanamori, Y., and Nakashima, N. (2001) A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation, *RNA 7*, 266-274.

9. Cannone, J., Subramanian, S., Schnare, M., Collett, J., D'Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Müller, K., Pande, N., Shang, Z., Yu, N., and Gutell, R. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs., *BMC Bioinformatics 3*, e2.

10. Khvorova, A., Lescoute, A., Westhof, E., and Jayasena, S. D. (2003) Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity, *Nat Struct Biol 10*, 708-712.

11. De la Pena, M., Gago, S., and Flores, R. (2003) Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity, *EMBO J 22*, 5561-5570.

12. Flor, P. J., Flanegan, J. B., and Cech, T. R. (1989) A conserved base pair within helix P4 of the Tetrahymena ribozyme helps to form the tertiary structure required for self-splicing, *EMBO J 8*, 3391-3399.

13. Murphy, F. L., and Cech, T. R. (1994) GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain, *J Mol Biol 236*, 49-63.

14. Wang, C., Le, S. Y., Ali, N., and Siddiqui, A. (1995) An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5' noncoding region, *RNA 1*, 526-537.

15. Barton, G. J. (2002) Barton, G.J. OC: a cluster analysis program. University of Dundee. Scotland, UK, 2002.

16. Kleywegt, G. J. J., T. A. (1994) A super position, *ESF/CCP4 Newsletter 31*, 9-14.

17. Boehringer, D., Thermann, R., Ostareck-Lederer, A., Lewis, J. D., and Stark, H. (2005) Structure of the hepatitis C virus IRES bound to the human 80S ribosome: remodeling of the HCV IRES, *Structure 13*, 1695-1706.

18. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis, *J Comput Chem 25*, 1605-1612.

19. Lukavsky, P. J., Kim, I., Otto, G. A., and Puglisi, J. D. (2003) Structure of HCV IRES domain II determined by NMR, *Nat Struct Biol 10*, 1033-1038.

20. Collier, A. J., Gallego, J., Klinck, R., Cole, P. T., Harris, S. J., Harrison, G. P., Aboul-Ela, F., Varani, G., and Walker, S. (2002) A conserved RNA structure within the HCV IRES eIF3-binding site, *Nat Struct Biol 9*, 375-380.

21. Lukavsky, P. J., Otto, G. A., Lancaster, A. M., Sarnow, P., and Puglisi, J. D. (2000) Structures of two RNA domains essential for hepatitis C virus internal ribosome entry site function, *Nat Struct Biol 7*, 1105-1110.

22. Kieft, J. S., Zhou, K., Grech, A., Jubin, R., and Doudna, J. A. (2002) Crystal structure of an RNA tertiary domain essential to HCV IRES-mediated translation initiation, *Nat Struct Biol 9*, 370-374.

23. Spahn, C. M., Kieft, J. S., Grassucci, R. A., Penczek, P. A., Zhou, K., Doudna, J. A., and Frank, J. (2001) Hepatitis C virus IRES RNA-induced changes in the conformation of the 40s ribosomal subunit, *Science 291*, 1959-1962.

24. Delano, W. The Pymol molecular graphics system, Delano Scientific, San Francisco.

25. Costantino, D. A., Pfingsten, J. S., Rambo, R. P., and Kieft, J. S. (2008) tRNA-mRNA mimicry drives translation initiation from a viral IRES, *Nat Struct Mol Biol 15*, 57-64.

26. Martick, M., and Scott, W. G. (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis, *Cell 126*, 309-320.

27. Westhof, E., Dumas, P., and Moras, D. (1988) Restrained Refinement of 2 Crystalline Forms of Yeast Aspartic-Acid and Phenylalanine Transfer-Rna Crystals, *Acta Crystal A44*, 112-123.

28. Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing, *Science 273*, 1678-1685.

29. Honda, M., Brown, E. A., and Lemon, S. M. (1996) Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA, *RNA 2*, 955-968.

RNA	Secondary structure	Tertiary interactions	Inferred pair-wise tertiary contacts	Crystal structure
CrPV	ref. (*1*)	DMS and kethoxal probing (*1*), mutational analysis (*8*)	U16-G47 A17-U46 G18-C45 G19-C44 U20-A43 A21-U42 G22-U41	3b31 (*25*)
HHR	(*2*)	Mutational analysis (*10, 11*)	A7-A27 A7-U30 C61-A27 C61-U30	2goz (*26*)
$tRNA^{asp}$	(*3*)	Chemical protection (*3*), sequence covariation (*9*)	U8-A14 A15-U47 G18-U54 U19-C55	2tra (*27*)
P546	(*4*)	Mutational analysis, hydroxyl radical protection (*13*)	A51-C121 A51-G148	1gid (*28*)
		Mutational analysis, DMS modification (*12*)	A81-C7 A81-110	
HCV-PK	(*29*)	Mutational analysis, chemical and enzymatic probing, thermodynamic calculations (*14*)	U72-A96 G73-U95 C74-G94 G75-C93 A76-U92 G77-C91	*none*

Table S1. Sources of secondary and tertiary structure information. All secondary and tertiary structure information was available prior to crystallographic structure determination. Note that the numbering of the residues is that of the sequences used in modeling (each modeled RNA beginning with position 1) and is not necessarily the same as the source references.

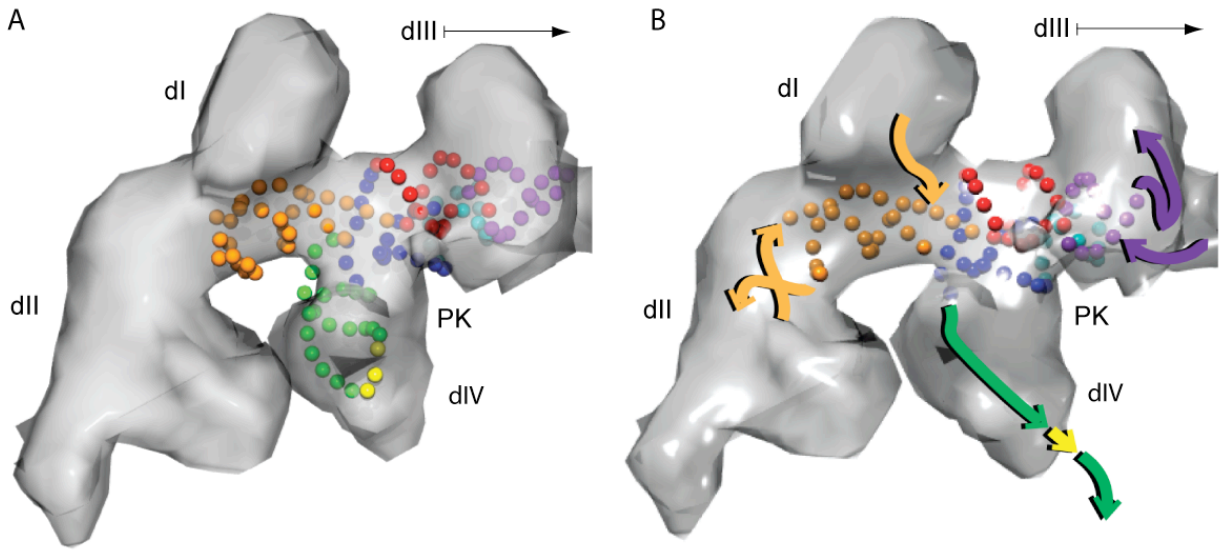

Figure S1. **HCV-PK model with domain IV in (A) base paired and (B) extended conformations fit into the cryo-EM density of the HCV IRES.** Arrows in panel B emphasize connections between the core pseudoknot and 4-way junction and other regions of the IRES.

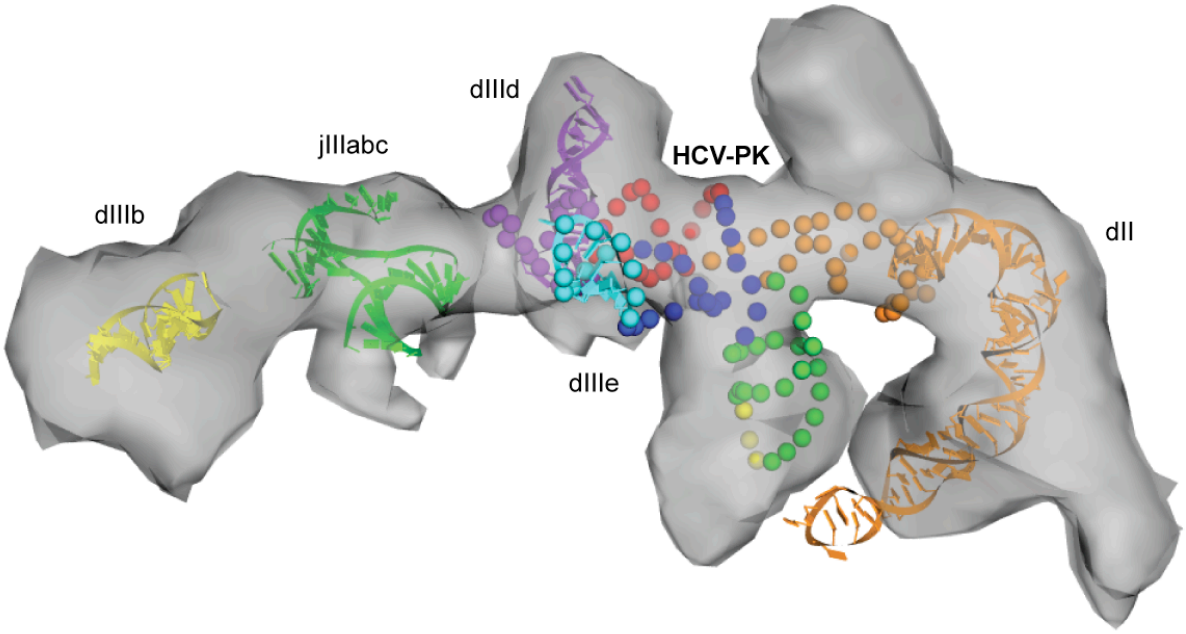

Figure S2. **HCV-PK model fit into the cryo-EM density of the HCV IRES with prior high-resolution structures.** The HCV-PK model is shown as spheres corresponding to phosphate pseudoatom positions, colored as in Figure 3. The cryo-EM density of the HCV IRES (in grey) is from the complex with human 80S ribosomes (*17*). High-resolution structures (*19*) are displayed using backbone cartoons.