Supporting Online Material

Materials and Methods

Protein expression and purification

Ten-subunit *S. cerevisiae* RNA polymerase II (Pol II) was purified essentially the same as previously described (*1*). DNA corresponding to *S. cerevisiae* TFIIB (residues 50-217) was cloned into the pTYB2 vector (New England Biolabs) using Nde I and Sma I restriction sites. Fusion protein was expressed in Rosetta 2(DE3) cells (Novagen) and purified by chitin affinity column. Following the cleavage of the affinity tag TFIIB (residues 50-217) was further purified by chromatography on SP sepharose ion exchange column and S200 Superdex gel filtration column.

Crystallization, data collection and structure refinement

RNA oligonucleotides (unmodified 3 nt – 9 nt, brominated 5 nt and 3'-deoxy 5 nt; see Table 1) were purchased from Dharmacon and were dissolved in ultrapure water as a 1mM stock. RNA dinucleotide (*i.e.* 2 nt; see Table 1) was purchased from IBA GmbH and dissolved in ultrapure water as a 100mM stock. Template and nontemplate DNA oligonucleotides (Table 1) were purchased from IDT and were dissolved in ultrapure water as a 2mM stock. RNA-DNA hybrids were assembled by annealing RNA, template DNA and nontemplate DNA with a molar ratio of 2:1:1 in an annealing buffer containing 20mM Tris-HCl, pH 8.0, 40mM KCl and 5mM $MgCl_2$.

To assemble Pol II – RNA – DNA ternary initiation complexes 10 mg/ml Pol II in a buffer containing Tris-HCl, pH 8.0, 40mM KCl, 5mM $MgCl_2$ and 5mM DTT was mixed with a 5-fold molar excess of the RNA-DNA hybrids, followed by the addition of a 10-fold molar excess of TFIIB (residues 50-217). Crystals of the ternary complexes were grown in a crystallization solution containing 390mM $(NH_4)_2HPO_4/NaH_2PO_4$, pH 5.9-6.1, 50mM dioxane, 9-11% PEG 6,000 and 15mM DTT, and were cryo-protected in a buffer containing 100mM Mes, pH 6.0, 50mM dioxane, 16% PEG 6,000, 350mM NaCl, 17% PEG 400 and 10mM DTT as described (*2*). The initiation complexes formed plate-like crystals, distinct from the crystals of the apo Pol II without nucleic acids, which were irregular and rod-like in shape. This difference provided a fast and reliable way to assess the capture of the short RNA-DNA hybrids by including various TFIIB constructs in the crystallization trials. In case of the RNA dinucleotide, the cryo-protectant supplemented with a final concentration of 5mM of the dinucleotide was used to soak overnight the crystals of a Pol II – DNA initiation complex prepared essentially as described above, except that no RNA was added when assembling the complex. In case of NTP soaking (*i.e.* ATP, GTP or 2'-iodo ATP, purchased from Jena Biosciences), the crystals of the corresponding initiation complexes were soaked overnight in 5mM NTP supplemented in the cryo-protectant.

Native, Br-SAD and I-SAD X-ray diffraction data were collected at the Stanford Synchrotron Radiation Laboratory and the Advanced Photon Source (Table 2). Raw data were processed with HKL2000 (*3*) or Mosflm (*4*) and Scala (*5*) or XDS (*6*). Structures were solved by molecular replacement with the use of Phaser (*7*), with the pol II coordinates from pdb entry 2NVQ as the search model. In case of Br-SAD and I-SAD data, anomalous signals were determined by SAD with molecular partial structure module in Phaser (*7*). Iterative model building were done with Coot (*8*) and the structural models were subject to simulated annealing with Phenix (*9*) followed by a final refinement with Buster-TNT (*10*). The downstream duplex was omitted from model building for all the initiation complexes. Although TFIIB (residues 50-217) was proved necessary for obtaining the crystals of the initiation complexes, no electron density could be attributed to it for any of the initiation complexes.

Molecular dynamics simulation of pol II initiation complexes

The dynamics of pol II initiation complexes with 3-nt, 4-nt, and 6-nt RNAs were investigated by all-atom molecular dynamics (MD) simulations with explicit water. The initial configurations for the systems with 4-nt and 6-nt RNAs were taken from the crystal structures obtained from the current study. Since only the 3'-end of the RNA-DNA hybrid is detectable in the X-ray structure for the 3-nt complex, the starting structure for MD simulation was modeled on the basis of the 4-nt complex structure by truncating its terminal nucleotide. Missing residues in the pol II structures were filled in by Segmod (*11*). In our MD simulations, simplified models of pol II containing the chain A and B were adapted since these are the only chains in direct contact with the RNA/DNA hybrids. Positions of the protein heavy atoms were restrained with a harmonic potential of 1000 kJ/mol^{-1}/nm^{-2}, while the RNA/DNA hybrids were allowed to move freely in order to investigate their dynamics. The pol II initiation complexes were placed in a water box, with a water layer extending at least 7 Å from the protein surface. One Mg^{2+} adjacent to the 3'-end of the RNA in the crystal structures was included, along with enough Cl^{-} ions to make the system electrically neutral. The entire systems contain 153,037, 153,015, and 158,960 atoms for 3-nt, 4-nt and 6-nt complexes respectively.

The ENCAD program was used for initial minimization to remove bad contacts (*12*), and subsequently the GROMACS simulation package was used to perform MD simulations (*13*). The AMBER03 force field was used for the protein, RNA, and DNA (*14*), and the TIP3P water model (*15*) was used for the explicit solvent. For electrostatic interactions, the Particle-Mesh Ewald method (*16*) was used with a short-range cutoff at 10 Å. For van der Waals interactions, a typical 9 Å cutoff was used and a switch function was applied to make the functions go smoothly to zero at 8 Å. A time step of 2.0 fs was used with lengths of the bonds involving hydrogen atoms constrained by the LINCS algorithm (*17*). A 10 Å neighbor list was updated every 10 time steps. Standard procedures, including a conjugate gradient minimization and a 50 ps NVT MD simulation with position restraints on the solute were used to equilibrate the solvated system. For each transcript the final configurations from equilibration were then used for four 1.5 ns production NVT simulations at 300K, with different initial velocities and Nose-Hoover thermostat for temperature coupling (*18*).

The terminal nucleotide for the 3-nt complex is highly mobile.

The root mean square fluctuation (RMSF) plots of ribonucleotides from the 3-nt RNA were compared with those from the 4-nt and 6-nt RNAs (**Fig. S3B**). RMSF ($RMSF = \sqrt{\frac{1}{T}\sum_{i=1}^{N}(x_i - \overline{x_i})^2}$) is a measure of average deviation of the particle positions. In our studies, RMSF for each nucleotide is computed using the center of mass of all nytrogen aotms on the base. The average is taken over conformations extracted from the last 1ns of the four 1.5ns simulations at an interval of 2ps. The structure was not fitted to the reference structure when computing the RMSF, so translational and rotational motions were also included. As shown in Fig. S3B, all the ribonucleotides of the 3-nt complex has a higher RMSF value compared to ribonucleotides at the same position of the 4-nt or 6-nt complex. More strikingly, the terminal ribonucleotide of the 3-nt complex at the i-3 position displayed much larger RMSF value than that of the 4-nt and 6-nt complex. These results indicate that the ribonucleotide at the i-3 position is much more flexible than systems containing longer RNA transcripts, which explains why only the 3'-end of the RNA-DNA hybrid is seen in the crystal structure. This unusual mobility of the ribonucleotide at the i-3 position of the 3-nt complex may due to the existence of a deep cavity in the hybrid-binding pocket of Rpb2 facing i-3 and a small fraction of i-4 ribonucleotide (**Fig. S3A**).

References

1. X. Liu, D. A. Bushnell, D. Wang, G. Calero, R. D. Kornberg, *Science* **327**, 206 (Jan 8, 2010).
2. D. Wang, D. A. Bushnell, K. D. Westover, C. D. Kaplan, R. D. Kornberg, *Cell* **127**, 941 (Dec 1, 2006).
3. Z. Otwinowski, M. Minor, *Methods Enzymol.* **276**, 307 (1997).
4. A. Leslie, *Crystallography* **26**, 27 (1992).
5. *Acta Crystallogr D Biol Crystallogr* **50**, 760 (Sep 1, 1994).
6. W. Kabsch, *Acta Crystallogr D Biol Crystallogr* **66**, 125 (Feb).
7. A. McCoy *et al.*, *Journal of Applied Crystallography* **40**, 658 (2007).
8. P. Emsley, K. Cowtan, *Acta Crystallographica Section D: Biological Crystallography* **60**, 2126 (2004).
9. P. D. Adams *et al.*, *Acta Crystallogr D Biol Crystallogr* **66**, 213 (Feb).
10. G. Bricogne *et al.*, BUSTER, version 2.8.0. Cambridge, United Kingdom: Global Phasing Ltd. (2009).
11. M. Levitt, *J. Mol. Biol.* **226**, 507 (1992).
12. M. Levitt, *J. Mol. Biol.* **168**, 595 (1983).
13. E. Lindahl, B. Hess, D. van der Spoel, *J. Mol. Mod.* **7**, 306 (2001).
14. Y. Duan *et al.*, *J. Comp. Chem.* **24**, 1999 (2003).
15. L. J. William, C. Jayaraman, D. M. Jeffry, W. I. Roger, L. K. Michael. (AIP, 1983), vol. 79, pp. 926-935.
16. P. Procacci, T. Darden, M. Marchi, *J. Phys. Chem.* **100**, 10464 (1996).
17. B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije., *J. Comput. Chem.* **18**, 1463 (1997).
18. W. Hoover, *Phys. Rev. A* **31**, 1695 (1985).

Fig. S1. F_o-F_c electron density map for the hybrids of the 8mer and 9mer transcribing complexes. Color scheme is the same as in Fig. 2.

Fig. S2. F_o-F_c electron density map for the hybrids of the 3mer transcribing complexes. Color scheme is the same as in Fig. 2. A structure model based on the 4-nt hybrid is placed in the map to indicate the location of the hybrid. Note the absence of the electron density at the i+1 site.

Fig. S3. Computational simulation studies of the initiation complexes (A) The RNA binding surface of Pol II is colored in gray with the pronounced surface cavity shown in blue. The 3rd RNA nucleotide is highlighted in a white dotted circle; (B)RMS fluctuations derived from the molecular dynamics simulations for a modeled 3-nt (based on the structure of the 4-nt hybrid), a 4-nt and a 6-nt hybrid.

Fig. S4. Overall view of the structural alignment of the 4-nt and 5-nt hybrids. The DNA templates for the 4-nt and 5-nt transcribing complexes are shown in blue and cyan and RNAs are in pink and red respectively.

Fig. S5. Structures of the 5-nt complexes with sequence variances in the hybrid. The $2F_o$-F_c electron density for these two hybrids are shown in (A) and (B).

Fig. S6. F_o-F_c electron density map for the initiation complexes with sequences in the hybrids displaying altered structural transition patterns. (A) A 5-nt hybrid adopts the canonical hybrid structure; (B) A 6-nt hybrid displays a distorted conformation; (C) A 7-nt hybrid with the sequence corresponding to that in (B) becomes the canonical hybrid structure.

Fig. S7. Pol II – hybrid interactions. RNA-Pol II (blue sticks) and DNA-Pol II (purple sticks) interactions are shown for the 6-nt hybrid. The lysine and arginine residues from Pol II that are located within 5Å of the phosphate backbones are displayed as sticks.

Movie S1. MD simulations of the 3-nt and 4-nt RNA Pol II initiation complexes are superimposed in the movie. Pol II is shown in the gray surface representation. The RNA chains of the 3-nt and 4-nt complexes are shown in green and red lines respectively. The i-3 ribonucleotides of both complexes are highlighted in the stick representation. The movie shows that the i-3 ribonucleotide of the 3-nt complex undergoes large fluctuations during the dynamics. This may be due to a deep cavity of the Pol II surface close to the i-3 ribonucleotide.

Table S1. DNA and RNA sequences

DNA	
Scaffold 1	3′-GAT GGC TAT TCG TC-5′ 5′-CTA CCG ATA AGC AGA CGA TCC TCT CGA TG-3′
Scaffold 1 Sequence variant 1	3′-GAT GGC TAT TCG TC-5′ 5′-CTA CCG ATA AGC AGA CGA TCG TCT CGA TG-3′
Scaffold 1 Sequence variant 2	3′-GAT GGC TAT TCG TC-5′ 5′-CTA CCG ATA AGC AGA CGA TGC TCT CGA TG-3′
Scaffold (5 nt (5Br-U))	3′-GAT GGC TAT TCG TC-5′ 5′-CTA CCG ATA AGC AGA CGA TCA CCT CGA TG-3′
RNA	
2 nt	5′-GG-3′
3 nt	5′-AGG-3′
4 nt	5′-GAGG-3′
5 nt	5′-AGAGG-3′
6 nt	5′-GAGAGG-3′
7 nt	5′-CGAGAGG-3′
8 nt	5′-TCGAGAGG-3′
9 nt	5′-ATCGAGAGG-3′
5 nt sequence variant 1	5′-AGACG-3′
5 nt sequence variant 2	5′-AGAGC-3′
5 nt (5Br-U)	5′-AGG(5Br-U)G-3′
5 nt 3’-deoxy	5′-AGAG(3′-deoxy-G)-3′
5 nt other sequence 1	5′-GGGAA-3′
7 nt other sequence 2	5′-AUCCUUA-3′

Table S2. Crystallographic data and structure statistics

Pol II-DNA-RNA Initiation Complexes, Native			
RNA	2 nt	3 nt	4 nt
Space group	C2	C2	C2
Unit cell dimension (Å)	165.4, 221.8, 193.8	158.1, 220.8, 191.9	161.0, 221.1, 192.9
(°)	90.0, 99.6, 90.0	90.0, 97.7, 90.0	90.0, 98.3, 90.0
Wavelength (Å)	0.979	0.979	0.979
Resolution (Å)	50 – 3.6	50 – 2.9	50 – 3.0
Unique reflections	79,458	143,811	132,760
Completeness (%) [a]	99.8 (99.3)	99.8 (98.8)	99.5 (97.9)
Redundancy	3.8 (3.4)	3.8 (3.4)	3.6 (3.2)
I/sigma	10.5 (1.8)	14.9 (1.5)	11.8 (1.2)
R_{merge} (%) [b]	13.9 (64.8)	10.1 (78.9)	10.9 (76.8)
R_{cryst} / R_{free} (%) [c, d]	19.0 / 23.4 [e]	19.4 / 23.0 [e]	18.8 / 23.5
RMS deviations			
Bond lengths	0.008	0.009	0.01
Bond angles	1.16	1.23	1.28
PDB accession	N / A	N / A	3RZO

Pol II-DNA-RNA Initiation Complexes, Native			
RNA	5 nt	6 nt	7 nt
Space group	C2	C2	C2
Unit cell dimension (Å)	157.5, 221.3, 192.2	159.3, 221.3, 192.4	167.9, 220.9, 194.6
(°)	90.0, 97.4, 90.0	90.0, 98.0, 90.0	90.0, 100.2, 90.0
Wavelength (Å)	0.979	0.979	0.979
Resolution (Å)	50 – 3.3	50 – 2.85	50 – 3.3
Unique reflections	97,374	154,876	104,957
Completeness (%)	99.4 (99.2)	99.9 (99.8)	99.9 (99.9)
Redundancy	3.8 (3.8)	3.8 (3.6)	3.8 (3.7)
I/sigma	7.9 (1.8)	13.6 (1.4)	12.6 (1.8)
R_{merge} (%)	17.3 (76.9)	11.0 (91.1)	13.2 (76.1)
R_{cryst} / R_{free} (%)	18.8 / 23.0	18.9 / 22.7	17.3 / 22.7
RMS deviations			
Bond lengths	0.008	0.01	0.01
Bond angles	1.14	1.32	1.35
PDB accession	3RZD	3S14	3S15

Pol II-DNA-RNA Initiation Complexes, Native		
RNA	8 nt	9 nt
Space group	C2	C2
Unit cell dimension (Å)	157.8, 221.1, 192.8	160.0, 220.9, 192.0
(°)	90.0, 97.6, 90.0	90.0, 98.2, 90.0
Wavelength (Å)	0.979	0.979
Resolution (Å)	30 – 3.25	30 – 3.2
Unique reflections	103,014	109,833
Completeness (%)	99.4 (96.6)	100.0 (100.0)
Redundancy	3.9 (3.8)	3.9 (3.8)
I/sigma	10.1 (1.4)	12.1 (1.9)
R_{merge} (%)	14.6 (81.1)	12.4 (69.1)
R_{cryst} / R_{free} (%)	18.3 / 22.7	18.8 / 23.6
RMS deviations		
Bond lengths	0.009	0.009
Bond angles	1.18	1.26
PDB accession	3S16	3S17

Pol II-DNA-RNA Initiation Complexes, Native		
RNA	5 nt sequence variant 1	5 nt sequence variant 2
Space group	C2	C2
Unit cell dimension (Å)	156.9, 220.7, 191.8	157.3, 221.1, 191.9
(°)	90.0, 97.5, 90.0	90.0, 97.5, 90.0
Wavelength (Å)	0.979	0.979
Resolution (Å)	30 – 3.13	50 – 3.1
Unique reflections	113,147	119,031
Completeness (%)	99.1 (94.3)	100.0 (100.0)
Redundancy	3.7 (3.1)	3.8 (3.7)
I/sigma	12.2 (1.6)	11.9 (1.8)
R_{merge} (%)	10.5 (50.6)	11.9 (72.9)
R_{cryst} / R_{free} (%)	18.3 (23.3)	19.0 (23.2)
RMS deviations		
Bond lengths	0.009	0.009
Bond angles	1.24	1.18
PDB accession	3S1M	3S1N

Pol II-DNA-RNA Initiation Complexes, Derivatized		
RNA	5 nt + 2’-iodo ATP	5 nt (5Br-U)
Space group	C2	C2
Unit cell dimension (Å)	165.5, 221.6, 194.1	158.3, 220.8, 192.2
(°)	90.0, 99.7, 90.0	90.0, 97,7, 90.0
Wavelength (Å)	1.2651	0.91499
Resolution (Å)	50 – 3.3	50 – 3.2
Unique reflections	100,015	107,380
Completeness (%)	96.5 (77.6)	98.8 (91.5)
Redundancy	6.8 (3.9)	7.0 (4.9)
I/sigma	16.3 (1.7)	11.0 (1.5)
R_{merge} (%)	11.4 (84.7)	15.3 (59.2)
R_{cryst} / R_{free} (%)	19.0 (24.0)	18.2 (23.0)
RMS deviations		
Bond lengths	0.009	0.009
Bond angles	1.22	1.20
Anomalous peak (σ) [f]	7.0	9.1
PDB accession	3S2H	3S2D

Pol II-DNA-RNA Initiation Complexes, Native		
RNA	5 nt 3’-deoxy + ATP	5 nt 3’-deoxy + GTP
Space group	C2	C2
Unit cell dimension (Å)	161.9, 220.5, 193.7	160.7, 221.3, 193.2
(°)	90.0, 98.5, 90.0	90.0, 98.3, 90.0
Wavelength (Å)	0.979	0.979
Resolution (Å)	30 – 3.3	30 – 3.2
Unique reflections	99,954	108,690
Completeness (%)	98.7 (94.9)	99.4 (95.5)
Redundancy	3.8 (3.8)	3.8 (3.7)
I/sigma	9.0 (1.6)	9.3 (1.7)
R_{merge} (%)	11.8 (100.2)	16.4 (107.1)
R_{cryst} / R_{free} (%)	17.6 (22.9)	17.7 (22.6)
RMS deviations		
Bond lengths	0.01	0.01
Bond angles	1.36	1.30
PDB accession	3S1Q	3S1R

[a] Values in parentheses are from the highest resolution shell.

[b] $R_{merge} = \Sigma|I-\langle I\rangle|/\Sigma\langle I\rangle$

[c] $R_{cryst} = \Sigma||F_o|-|F_c||/\Sigma|F_o|$

[d] $R_{free} = \Sigma_T||F_o|-|F_c||/\Sigma_T|F_o|$, where T is a test data set of 5% of the total reflections randomly chosen and set aside before refinement

[e] This statistics corresponds to refinements using apo Pol II

[f] Anomalous peak heights are derived from the anomalous Fourier map generated using phases from Phaser ‘SAD with molecular replacement partial structure’

Fig. S1

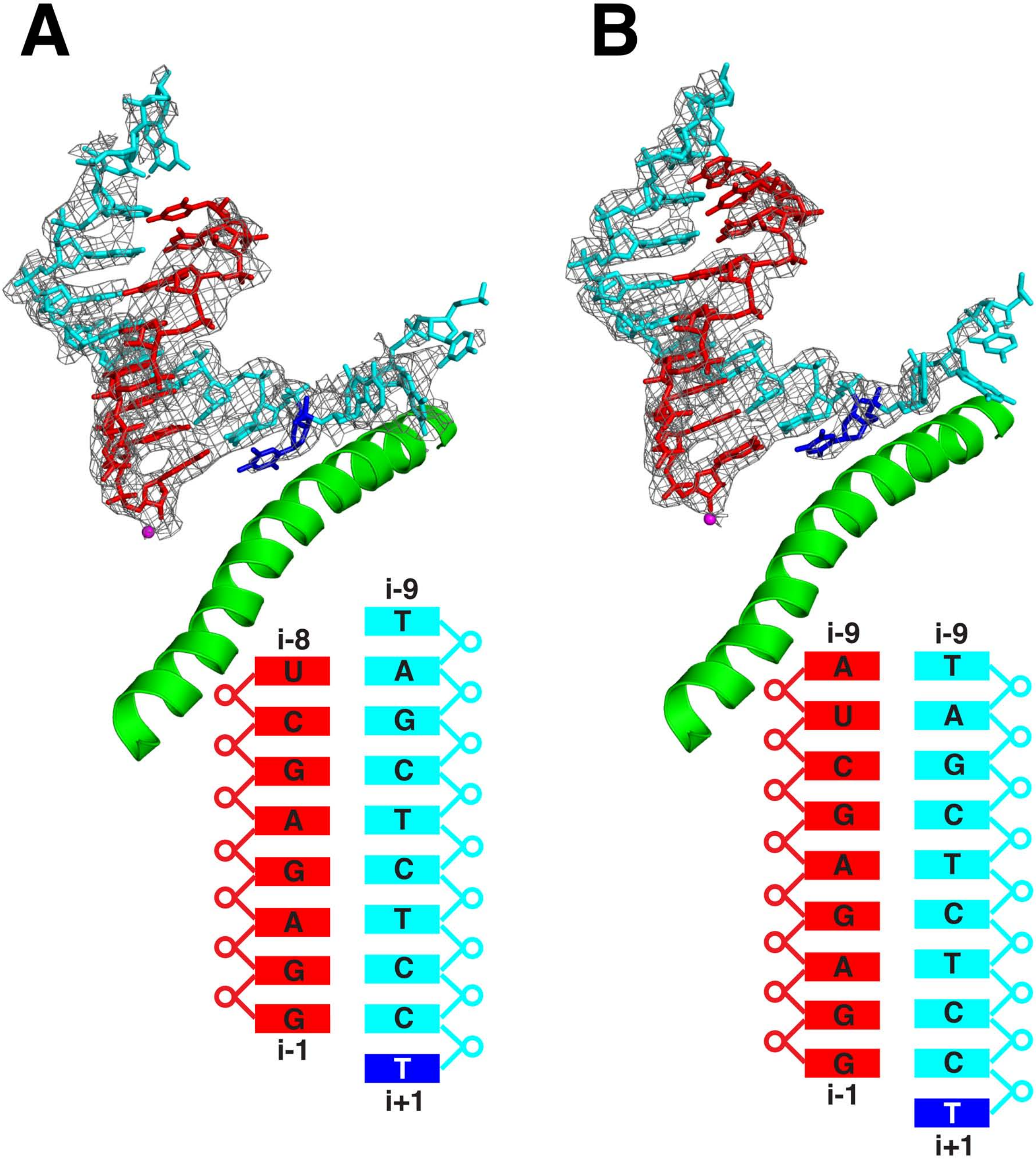

A
i-9
i-8
U
T
C
A
G
G
A
C
G
T
A
C
G
T
A
C
G
C
i-1
T
i+1
B
i-9
i-9
A
T
U
A
C
G
G
C
A
T
G
C
A
T
G
C
G
C
i-1
T
i+1

Fig. S2

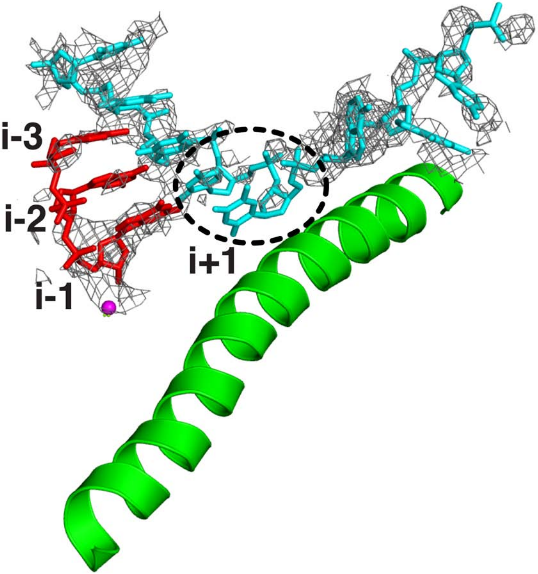


Fig. S3

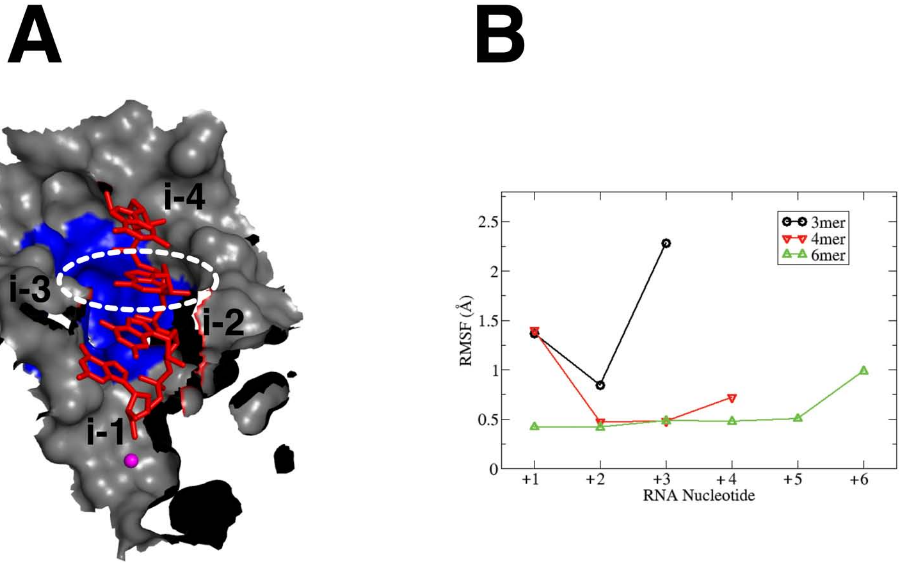

A
B
i-4
i-3
i-2
i-1
3mer
4mer
6mer
RMSF (Å)
RNA Nucleotide
0
0.5
1
1.5
2
2.5
+1
+2
+3
+4
+5
+6

Fig. S4

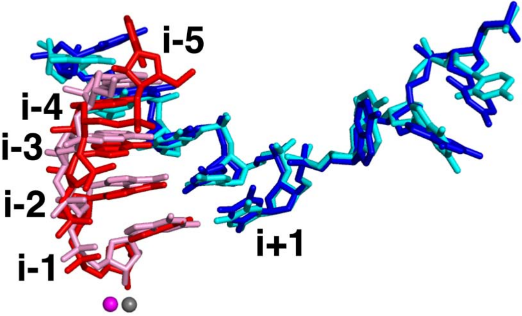


Fig. S5

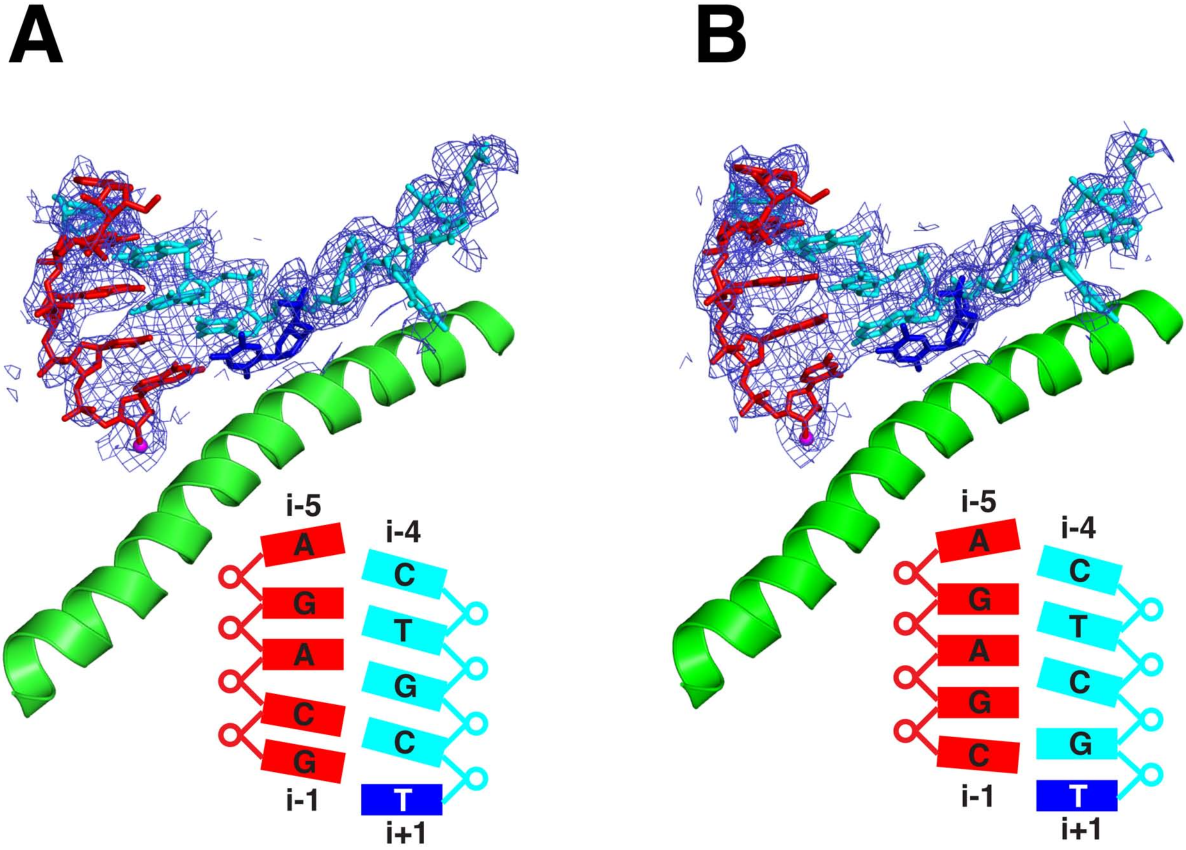


Fig. S6

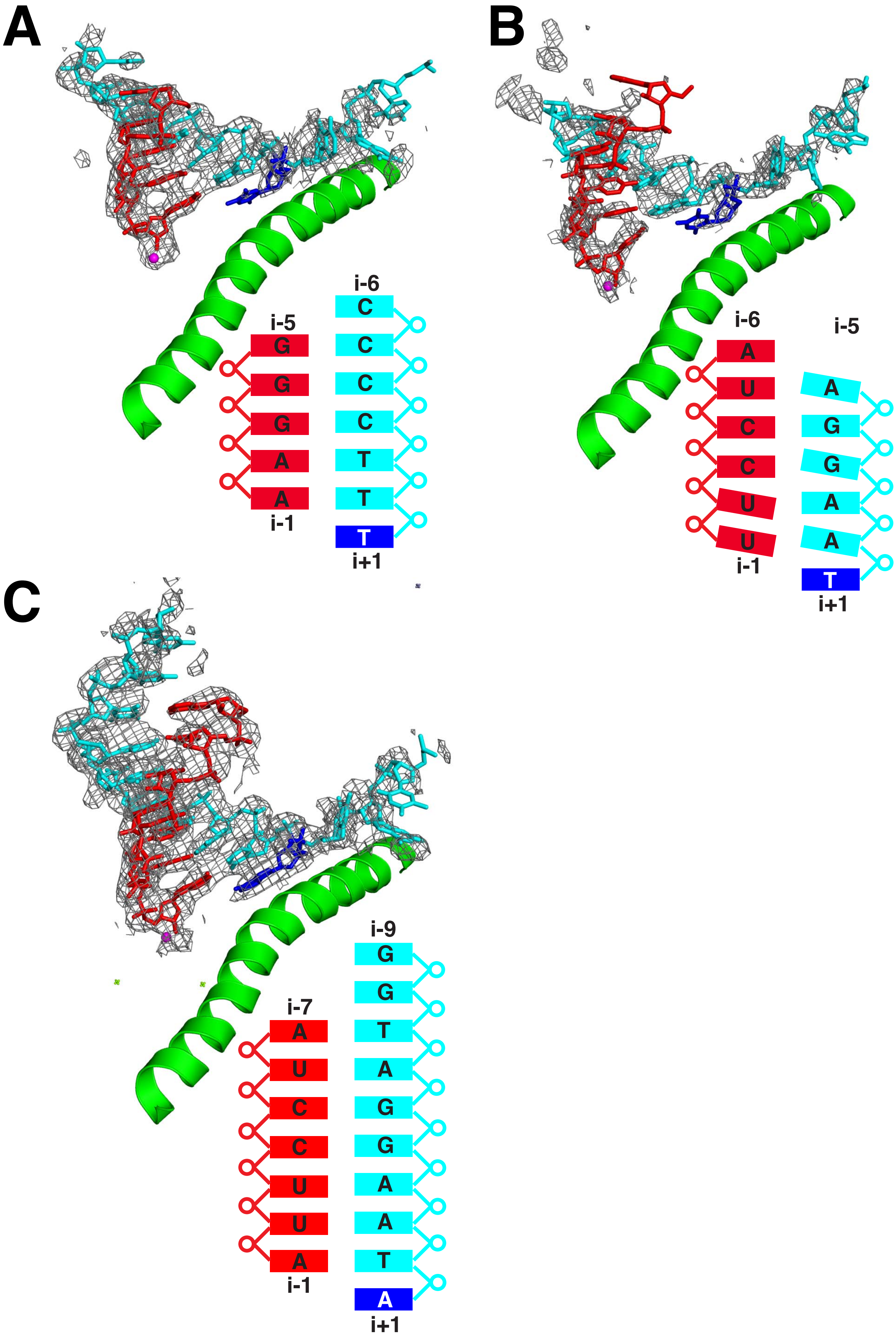
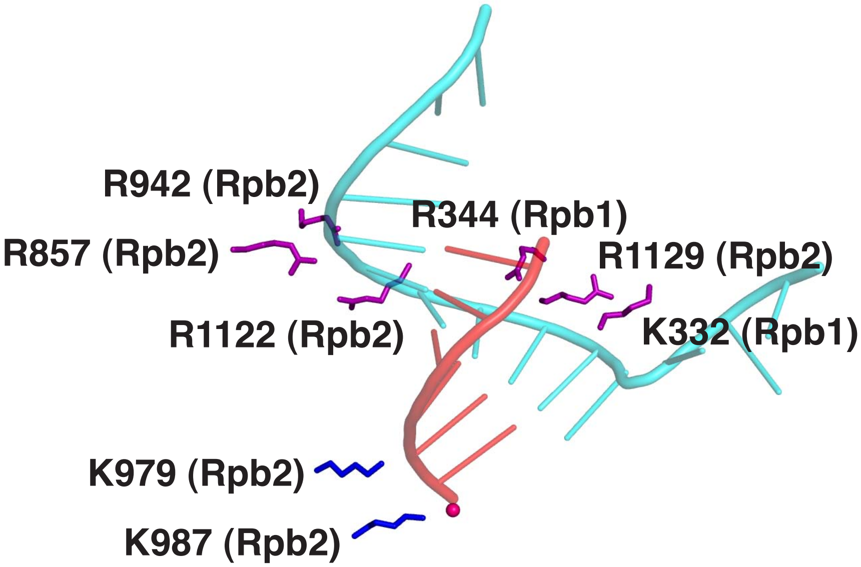

A
i-6
C
C
C
C
T
T
T
i+1
i-5
G
G
G
A
A
i-1
B
i-6
A
U
C
C
U
U
i-1
i-5
A
G
G
A
A
T
i+1
C
i-9
G
G
T
A
G
G
A
A
T
A
i+1
i-7
A
U
C
C
U
U
A
i-1

Fig. S7


R942 (Rpb2)
R344 (Rpb1)
R857 (Rpb2)
R1129 (Rpb2)
R1122 (Rpb2)
K332 (Rpb1)
K979 (Rpb2)
K987 (Rpb2)