Insight into the mechanism of nonenzymatic RNA primer extension from the structure of an RNA-GpppG complex

Wen Zhang,a,b,c,d Chun Pong Tan,a,b,c,e Travis Walton,a,b,c,d Albert C. Fahrenbach,a,b,c,d,f Gabriel Birран,g and Jack W. Szostak,a,b,c,d,e,f,1

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved June 6, 2017 (received for review March 9, 2017)

The nonenzymatic copying of RNA templates with imidazole-activated nucleotides is a well-studied model for the emergence of RNA self-replication during the origin of life. We have recently discovered that this reaction can proceed through the formation of an imidazolium-bridged dinucleotide intermediate that reacts rapidly with the primer. To gain insight into the relationship between the structure of this intermediate and its reactivity, we cocrystallized an RNA primer–template complex with a close analog of the intermediate, the triphosphate-bridged guanosine dinucleotide GpppG, and solved a high-resolution X-ray structure of the complex. The structure shows that GpppG binds the RNA template through two Watson–Crick base pairs, with the primer 3′-hydroxyl oriented to attack the 5′-phosphate of the adjacent G residue. Thus, the GpppG structure suggests that the bound imidazolium-bridged dinucleotide intermediate would be preorganized to react with the primer by in-line 5′,2′ substitution. The structures of bound GppG and GpppG suggest that the length and flexibility of the 5′-5′ linkage are important for optimal preorganization of the complex, whereas the position of the 5′-phosphate of bound pGpG explains the slow rate of oligonucleotide ligation reactions. Our studies provide a structural interpretation for the observed reactivity of the imidazolium-bridged dinucleotide intermediate in nonenzymatic RNA primer extension.

RNA self-replication | diguanosine dinucleotide | crystal structure | origin of life

1To whom correspondence should be addressed. Email: szostak@molbio.mgh.harvard.edu.

Significance

Rudimentary mechanisms of genome replication are essential for the earliest RNA-based cellular life, yet it is unknown how RNA or related polymers could have replicated nonenzymatically. For decades, 2-methylimidazole-activated GMP (2-MeImpG) has been used as a model substrate. We recently showed that two 2-MeImpG monomers react to form an imidazolium-bridged dinucleotide, which then reacts rapidly with the RNA primer. To explore this mechanism, we cocrystallized an RNA primer–template complex with several 5′-5′-linked analogs of the imidazolium-bridged intermediate. The closest analog, GpppG, binds to RNA in a conformation that explains the high reactivity of the imidazolium-bridged intermediate, whereas the structures of other dinucleotide ligands appear less favorable. Our study provides insight into the fundamental mechanism of nonenzymatic RNA self-replication.

Author contributions: W.Z., C.P.T., and J.W.S. designed research; W.Z., C.P.T., and T.W. performed research; A.C.F. and G.B. contributed new reagents/analytic tools; W.Z., C.P.T., T.W., A.C.F., G.B., and J.W.S. analyzed data; and W.Z., C.P.T., and J.W.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Frequently available online through the PNAS open access option.

Data deposition: Crystallography, atomic coordinates, and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes SUEE, SUED, SUEG, and SUEF).

*Howard Hughes Medical Institute, Massachusetts General Hospital, Boston, MA 02114; †Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114; ‡Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA 02114; †Department of Genetics, Harvard Medical School, Boston, MA 02115; ′Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; †Division of Experimental Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215

© 2017 by The National Academy of Sciences

intermediate (Fig. 1B). As yet, there is no structural evidence as to how the dinucleotide intermediate binds the RNA and whether the conformation would be favorable for nucleophilic attack.

Despite these recent studies, relatively few mechanistic investigations have been based on crystallographic approaches. We have previously used crystallographic approaches to investigate various aspects of nonenzymatic RNA polymerization, including the effect of 2′-5′ phosphodiester linkages and 2-thio-U substitution on RNA structure (19–21). To study either enzymatic or nonenzymatic replication reactions, chemically stable analogs of substrates are typically necessary. For example, the nonhydrolyzable α-β imido ATP analogs have been intensively applied to structural studies of the mechanism of catalysis by DNA polymerase (22). In the J.W.S. laboratory, the nonhydrolyzable guanosine 5′-(3-methyl-1H-pyrazol-4-yl)phosphonate (PZG) was designed and synthesized to mimic the activated monomer 2-MeImpG (23). Our crystal structures of RNA-PZG complexes revealed both Watson–Crick and, surprisingly, noncanonical base pairs, suggesting that mismatched template–monomer base pairing may be more common than expected. More significantly, the phosphate and leaving group portions of these structures were always disordered. The absence of any detectable interaction between the leaving group analogs of adjacent monomers suggested that the catalytic effect of a downstream activated nucleotide might not be due to a noncovalent interaction and raised the question of whether the imidazolium-bridged dinucleotide intermediate might exhibit a more ordered structure that could help to explain its high reactivity.

To address the mechanism of nonenzymatic RNA polymerization, we used X-ray crystallography to provide high-resolution structures of RNAs complexes with stable analogs of the reactive intermediate. RSN2 reaction and primer extension. (Sn pound, imidazolium-bridged intermediate. The readily available complexes of RNAs complexed with stable analogs of the reactive intermediates, we used X-ray crystallography to provide high-resolution structures, the RNA cocrystallized with the GpppG ligand with hexagonal symmetry, as in our previous RNA-GMP and RNA-PZG complexes (23). We determined the structure to 1.9-Å resolution by molecular replacement, with an overall B-factor of 32.42. The space group is P3121, and there is one RNA duplex with two bound GpppG molecules per asymmetric unit. At each end of the RNA duplex, the mCmC overhang serves as the binding site for G-G dinucleotides because the mC residues form Watson–Crick base pairs with G (23, 27). The RNA coocrystallized with the GpppG ligand with hexagonal symmetry, as in our previous RNA-GMP and RNA-PZG complexes (23).

Results

Overall Structural Features of the RNA-GpppG Complex. We co-crystallized a series of guanosine dinucleotides with the RNA 5′-mCmCmCGACUAAAG-UCG-3′ (SI Appendix provides experimental details). The first three nucleotides in bold are 5-methylcytidine LNA residues, designed to favor and rigidify the A-form strand conformation (26) and thereby facilitate RNA crystallization. At each end of the RNA duplex, the mCmC overhang serves as the binding site for G-G dinucleotides because the mC residues form Watson–Crick base pairs with G (23, 27). The RNA coocrystallized with the GpppG ligand with hexagonal symmetry, as in our previous RNA-GMP and RNA-PZG complexes (23). We determined the structure to 1.9-Å resolution by molecular replacement, with an overall B-factor of 32.42. The space group is P3121, and there is one RNA duplex with two bound GpppG molecules per asymmetric unit. At each end of the RNA duplex, the overhanging mCmC binding sites are fully occupied through Watson–Crick base pairs with GpppG (Fig. 2 A and B). As in our previously determined RNA–monomer complex structures, the RNA double helices are A-form, all sugars are in the C3′-endo conformation, and the duplexes slip-stack on each other to form extended columns. Groups of three RNA duplex-ligand complexes form triangular prisms, and the central channel accommodates at least three water molecules that bridge the neighboring duplexes (Fig. 2C). Two symmetry-related water molecules form three 2.5-Å H-bond contacts with the three surrounding duplexes via the 2′-hydroxyls of their G4 residues (SI Appendix, Fig. S2A).

The other water (and a possible fourth water molecule at a symmetry-related position, but with very weak density) appears to engage in three 3.1-Å hydrogen bonds with the pro-Sο nonbonded oxygen atoms of three G4–A5 phosphodiester linkages. However, this may actually be a time-averaged view of a water molecule that at any given moment is acting as an H-bond donor to two nonbridging phosphate oxygens. In addition, Mg2+ ions link the three adjacent complexes by coordinating with both the 2′- and 3′-hydroxyl groups of the guanosine at the primer +1 position of three adjacent GpppG ligands, forming a total of six ∼2.4-Å electrostatic interactions. (Fig. 2D).

Structure of GpppG Bound to RNA. A GpppG ligand is Watson–Crick base paired to the 5′-mCmC overhang at each end of the RNA duplex; the six hydrogen bond contact distances range from 2.8 Å to 3.0 Å (Fig. 2 E and F). The entire GpppG...
The molecule is well-ordered. The two guanine nucleobases are co-planar and stacked with the upstream primer and the neighboring duplex with interplanar distances of ∼3.3 Å. Both ribose sugars of the GpppG are in the C3'-endo A-form conformation, consistent with our previous observation of a C3'-endo sugar pucker when activated guanosine ribonucleotides bind to a template in solution (28). Moreover, the GpppG triphosphate linkage is well-ordered due to a Mg2+ ion that is coordinated with three nonbridging oxygen atoms, one from each phosphate of GpppG. The three electrostatic interaction distances are 2.3 Å, 2.5 Å, and 3.0 Å. This coordination with Mg2+ results in the triphosphate bridge of GpppG being buckled in a well-defined manner (Fig. 2E and F).

Because GpppG was chosen as a close analog of the imidazolium-bridged dinucleotide Gp-Im-pG, we were interested in whether GpppG is bound in a conformation consistent with the observed high reactivity of the Gp-Im-pG intermediate with the primer. The distance between the primer 3'-hydroxyl and the phosphorus atom of the closest phosphate of GpppG is 4.1 Å, and the angle between the 3'-OH and the bridging P-O bond of GpppG is 126° (Table 1). We then asked whether the actual imidazolium-bridged intermediate could potentially adopt a similar conformation to the observed conformation of GpppG when bound to the RNA primer–template. We constructed a 2-aminoimidazolium–bridged diguanosine model (Gp-NH2Im-pG) and applied it in the restrained refinement in place of GpppG (29). We discovered that the Gp-NH2Im-pG ligand fit reasonably well to the electron density of GpppG (Fig. 2G). The observed electron density fits both the nucleobases and sugars of Gp-NH2Im-pG very well (B-factors: nucleobase, 27.5; sugar, 36.4), and the Gp-NH2Im-pG molecule formed two Watson–Crick base pairs with the template in the same manner as GpppG. The B-factors of the two phosphorus atoms in Gp-NH2Im-pG (37.6 and 45.6) were comparable with the corresponding atoms in GpppG (31.6 and 43.4), and the distance between the P1 and P3 atoms in Gp-NH2Im-pG is only marginally longer than that in GpppG (4.8 Å vs. 4.5 Å). The fact that Gp-NH2Im-pG can be modeled to fit the density corresponding to GpppG demonstrates the potential structural similarity between the imidazolium-bridged diguanosine intermediate and the stable GpppG analog, when bound to an RNA primer–template complex.

In order for the proposed dinucleotide intermediate in non-enzymatic RNA polymerization to bind and react with the primer, it must compete with the more abundant free monomer for template occupancy. We therefore set out to measure the strength of GpppG binding to the RNA template relative to GMP. NMR methods have been used to measure the affinity of GMP for RNA primer–templates (10, 12–14). Here we used the same methods and RNA primer–template system (10) to determine
the affinity of GpppG for an RNA template overhang consisting of two consecutive cytidines. In the experiment, the concentrated GpppG solution was titrated into the RNA primer-template complex solution, and the binding constant was then calculated from the change in the chemical shift of the imino proton of the primer 3′-G vs. the concentration of GpppG (SI Appendix provides detailed procedures). Strikingly, the observed $K_d$ of GpppG is $\sim$0.2 mM, which is $\sim$100-fold greater than our previously determined measurement of the affinity of GMP, which has a $K_d$ of $\sim$20 mM (Fig. 2 H and I). The stronger binding of GpppG most likely reflects its two Watson–Crick base pairs with the template. We then asked whether the aminomimidazol-bridged diguanosine intermediate Gp-NH2Im-pG would have a similar affinity to the same RNA substrate. To address this question, we isolated Gp-NH2Im-pG in $\sim$80% purity, with 2-AmImpG as impurity, and used it as the substrate in a primer extension reaction using the same primer–template complex as for the $K_d$ measurement of GpppG. Interestingly, the $K_m$ is $\sim$0.6 mM, which is comparable to the $K_d$ of the GpppG analog, given the different conditions required for each assay (SI Appendix provides details). This affinity measurement suggests that the Gp-NH2Im-pG intermediate and the structural analog GpppG may compete effectively with monomers for binding to the template during nonenzymatic RNA polymerization reactions.

G(5′)ppp(5′)G Binds to RNA Template Through Two Different Motifs. To understand the properties of the 5′-5′ linkage that make the imidazolium-bridged dinucleotide an appropriate intermediate, we cocystallized the same RNA sequence with other guanosine dinucleotides that have different lengths or types of bridging linkages. Crystals of all the RNA–dinucleotide complexes grew with hexagonal symmetry (SI Appendix, Tables S2 and S3). The overall structures were similar to that of the RNA-GpppG complex, with the same molecular packing patterns, including the slip-stacked RNA double helices, the triangular-prism structure formed by groups of three RNA–ligand complexes, and the binding of the dinucleotide ligand with the mCmC binding sites at the RNA terminus. We first examined diguanosine-5′-5′-diphosphate (GpppG), which is known to form in trace amounts in activated monomer solutions as a result of the reaction of GMP, formed by hydrolysis, with an activated monomer. Complementary templates can catalyze the synthesis of pyrophosphate-linked dinucleotides from activated monomers (30). These pyrophosphate-bridged dinucleotides likely inhibit nonenzymatic RNA polymerization, presumably through competitive binding of the RNA template.

The mode of binding and the rigidity of the pyrophosphate linkage of GpppG are subtly different from that of GpppG. At one end, GpppG forms two Watson–Crick base pairs with the templating mCs, but the pyrophosphate linkage and sugars are disordered (Fig. 3B), making it difficult to define the geometry of the pyrophosphate and the conformation of the GpppG sugars. Remarkably, GpppG binds to the template in a distinctly different manner at the other end. The guanosine adjacent to the primer is Watson–Crick base paired with the template mC through three hydrogen bonds as expected. However, the second guanosine forms a noncanonical G:C base pair with two hydrogen bonds: a weak 3.5-Å H-bond between the guanine N3 and the exocyclic amine of the mC and a second 2.9-Å H-bond between the exocyclic amine of the guanine and the N3 of the mC. We previously observed the same type of noncanonical G:C base pair in our RNA-GMP and RNA-PZG structures, suggesting that this structure could play an important role in modulating the efficiency and fidelity of nonenzymatic RNA replication (23). This GpppG is well-ordered overall, including nucleobases, sugars, and the pyrophosphate linkage. The first ribose is in the C3′-endo conformation, whereas the second one is in the C2′-endo conformation. The distance between the 3′-hydroxyl group of the primer and the phosphorus center of pyrophosphate linkage is 4.8 Å (Table 1 and Fig. 3C), significantly longer than for the GpppG complex. Although the poor solubility of GpppG prevented us from measuring the affinity of GpppG for the RNA primer–template, the overall structure of the RNA-GpppG complex suggests that GpppG likely binds the RNA template with a similar affinity as GpppG.

Tetraphosphate of G(5)ppp(5)G Bound to an RNA Template Is Disordered. We then cocystallized the same RNA duplex with diguanosine-5′,5′-tetraphosphate (GpppGG), in which the linkage between the two nucleosides is longer and possibly more flexible than that of GpppG. In the complex structure, GpppGG binds to the RNA through two well-ordered G:C Watson–Crick base pairs at both ends, but the tetraphosphate linkage is disordered. The electron density associated with the tetraphosphate linkage (Fig. 4B) is significantly larger and more globular than that of GpppG and can be modeled as two GpppGG ligands with different conformations (each with an occupancy of 0.5). The distance between the 3′-hydroxyl group of the primer and the

---

Table 1. Crystallographic structure features

<table>
<thead>
<tr>
<th>RNA–ligand complex</th>
<th>5′-5′ linkage</th>
<th>Binding motifs</th>
<th>3′-O-P distance, Å</th>
<th>3′-O–P–O angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-GpppG</td>
<td>Triphosphate</td>
<td>Watson–Crick</td>
<td>4.1</td>
<td>126°</td>
</tr>
<tr>
<td>RNA-Gppp</td>
<td>Pyrophosphate</td>
<td>Watson–Crick/noncanonical</td>
<td>4.8</td>
<td>170°/ND</td>
</tr>
<tr>
<td>RNA-GppppG</td>
<td>Tetraphosphate</td>
<td>Watson–Crick</td>
<td>$\sim$4.4</td>
<td>ND</td>
</tr>
<tr>
<td>RNA-pGpG</td>
<td>Monophosphate (3′-5′)</td>
<td>Watson–Crick</td>
<td>4.9</td>
<td>ND</td>
</tr>
</tbody>
</table>

ND, not detectable.

---

Fig. 3. Structure of the RNA-GpppG complex. RNA and ligand structures and $F_o - F_c$ omit map are represented as mentioned above. (A) Chemical structure of GpppGG ligand. (B) At one end of RNA duplex, GpppG forms two Watson–Crick base pairs with template, but the diphosphate linkage and sugar moieties are disordered. (C) At the other end of the duplex, GpppG forms two different base pairs with the template. The diphosphate linkage is highly ordered.
adjacent phosphorus center is about 4.4 Å. The disordered structure of the tetraphosphate bridge suggests that any bridge with similar length and flexibility is unlikely to be optimal for the primer extension reaction due to unfavorable reaction angles for in-line attack by the primer.

**pGpG Binds the RNA Template Through Watson–Crick Base Pairs.** We cocrystallized the RNA with the 5'-phosphorylated GG dimer pGpG to compare this 3'-5' linked oligonucleotide with the previously studied 5'-5' bridged dinucleotides. At both ends of the RNA duplex, the pGpG dimer binds to the mCmcC binding sites through two Watson–Crick base pairs. However, the two sugars are partially disordered, making it impossible to define the ribose conformation. The 5'-phosphate of the dimer is displaced toward the major groove, and the distance between the 3'-hydroxyl of the primer and the phosphorus atom of the phosphate is 4.9 Å (Fig. 4D). This distance is the longest of all the complex structures and is consistent with the previous observation that primer extension by ligation is much slower than by polymerization with activated monomers in the presence of an activated downstream helper primer or oligonucleotide (16).

**Discussion**

Understanding the mechanism of template-directed non-enzymatic primer extension may lead to improved and/or more prebiotically realistic ways of driving RNA replication without enzymes. We have recently shown that nonenzymatic primer extension with 2-methylimidazole–activated nucleotides (5) can proceed via a two-step process, in which two activated monomers first react with each other to form an imidazolium-bridged di-nucleotide intermediate (18). Once formed and bound to the template, this intermediate reacts rapidly with the primer. This surprising reaction mechanism and the unusual structure of the intermediate raise several important questions. Here we have focused on the interaction of the intermediate with the template: Can the intermediate bind to the template without Watson–Crick base pairing of both of its nucleobases? Once bound, is there some aspect of its conformation that would favor reaction with the primer? To address these questions, we have taken advantage of the structural similarity between the triphosphate-bridged dinucleotide, GpppG, and the imidazolium-bridged diguanosine intermediate, Gp-Imp-p. The 2-methylimidazole/2-aminoimidazole–bridged dinucleotides are not stable enough to directly cocrystallize with RNA, but the triphosphate analog is quite stable. As a result, we were able to obtain a high-resolution crystal structure of a close analog to the true reaction intermediate bound to a RNA primer–template complex.

The crystal structure of the RNA-GpppG complex shows that both G nucleotides interact with the template via Watson–Crick base pairing. We also showed that Gp-NH2-Im-pG could be modeled to fit the electron density of GpppG. Based on the structural similarities of GpppG and the imidazolium-bridged intermediate, it seems highly likely that the true intermediate also binds to the template through two Watson–Crick base pairs. The formation of two Watson–Crick base pairs leads to a much higher affinity of GpppG than GMP for a CC template (Kd of ∼0.2 mM vs. ∼20 mM, respectively). Additionally, the Kd of purified Gp-NH2-Im-pG in primer extension (0.6 mM) suggests comparable affinity of the imidazolium-bridged intermediate.

The high affinity of the imidazolium-bridged intermediate helps to explain its effectiveness in primer extension reactions because even if only a small fraction of monomer is converted to intermediate, the intermediate would still be able to bind to the primer–template complex in the presence of a large excess of activated monomer. Interestingly, the GppG dinucleotide, which has a shorter linker between the two G residues, binds to the template via two Watson–Crick base pairs at one end of the RNA duplex, but at the other end by one Watson–Crick and one noncanonical base pair, suggesting that there is some strain involved in folding GppG into the conformation necessary to allow both Gs to Watson–Crick pair with the template.

Once the dinucleotide, GppG, is bound to RNA, its overall structure, including the sugars and the triphosphate linker, becomes highly ordered. In the RNA-GpppG complex, the triphosphate linkage is structured due to a Mg2+ ion that coordinates with one nonbridging oxygen from each phosphate. As a result, the local structure is preorganized for SN2 reaction with the primer 3'-hydroxyl, which in the case of the imidazolium-bridged intermediate would result in primer extension by one nucleotide. In the RNA-GpppG structure, the distance between the primer 3'-OH and the adjacent phosphate is 4.1 Å, and the 3'-O-P-O angle is 126°. The distance and angle seen here are comparable to the 4.1 Å and 107° for the Sn2 reaction catalyzed by the eukaryotic RNA polymerase II during transcription initiation, as seen in the corresponding crystal structure (31). Assuming that the imidazolium-bridged intermediate adopts a similar conformation, it is clear that further adjustments in the distance and angle of attack would have to occur before any reaction could take place, suggesting a possible role for divalent metal ion catalysis. Nevertheless, comparison with the structures of dinucleotides with different linkages (either a shorter pyrophosphate linkage or a longer tetraphosphate linkage) revealed greater 3'-O to P distances and/or greater disorder of the phosphate. Therefore, if the imidazolium-bridged intermediate adopts a conformation similar to that of GpppG under primer extension conditions, it will be partially preorganized for Sn2 attack by the primer hydroxyl. One caveat is that the intermediate generated from 2-methylimidazole–activated monomers could not bind Mg2+ in the same manner at the GpppG triphosphate, and the p-2MeIm-p bridge might be less ordered and therefore less optimally reactive. However, 2-aminoimidazole–activated monomers, which lead to 10-100 times faster primer extension than 2-methylimidazole–activated monomers (6), would generate a 2-aminoimidazolium–bridged intermediate. We propose that the 2-amino group of the imidazole could hydrogen bond to the nonbridging oxygens of both flanking phosphates, potentially resulting in a highly ordered structure that is preorganized for nucleophilic attack by the primer hydroxyl. It is also of note that when the dinucleotide pGpG is bound to the same primer–template complex, the primer 3'-O to P distance is 4.9 Å, which is significantly greater than the distance for any of the 5'-5' linked dinucleotides. This long distance may contribute to the slow rate of oligonucleotide
ligation reactions, compared with primer extension with activated monomers.

Finally, we note that the 5′-5′-linked dinucleotides with pyrophosphoryl groups are readily formed by attack of the 5′-phosphate of an unactivated monomer on the activated phosphate of a second monomer (30, 32). Similarly, attack of the β-phosphate of GDP on the phosphoanion of an activated GMP would generate GpppG. The tight binding of GpppG to a CC template suggests that GpppG might be an ideal primer for the initiation of template copying by primer extension. The resulting RNAs would begin with a 5′-cap-like GpppG moiety, suggesting a potential evolutionary origin for the eukaryotic mRNA 5′-cap structure.

In summary, our structural studies of template-bound GpppG support the model that the structurally similar imidazolium-bridged intermediate binds the template tightly through two Watson–Crick base pairs and that the conformational constraint imposed by the intercalative nucleotide bridge helps to preorganize the bound complex for in-line nucleophilic attack by the primer 3′-hydroxyl.

Materials and Methods

The oligonucleotide used for crystallography was custom-synthesized by Exiqon, Inc. Oligonucleotides for affinity measurements were prepared by solid-phase synthesis. Data were collected at the SIBYLs beamlines 8.2.1 and 8.2.2 at Lawrence Berkeley National Laboratory. Datasets were processed using HKL2000 and DENSITY SCALEPACK (33). All structures were solved by molecular replacement. The refinement protocol includes simulated annealing, positional refinement, restrained B-factor refinement, and bulk solvent correction (34). The topologies and parameters for m(CCC), GpppG(GP3), GpppG(GP2), and dinucleotide intermediate (GIM) were constructed and applied. Detailed experimental protocols are provided in SI Appendix.

Acknowledgments. We thank Drs. Li Li, Daniel Duzdevich, and Anna Wang for helpful discussions and insightful commentaries on the manuscript. This research used beamlines 8.2.1 and 8.2.2 of the Advanced Light Source of Lawrence Berkeley National Laboratory, which is a Department of Energy Office of Science User Facility under contract DE-AC02-05CH11231. J.W.S. is an investigator of the Howard Hughes Medical Institute. This work was supported in part by grant 290363 from the Simons Foundation (to J.W.S.) and by Grant CHE-1607034 from the NSF (to J.W.S.). A.C.F. was supported by a Research Fellowship from the Earth–Life Science Institute at the Tokyo Institute of Technology.
Insight into the mechanism of nonenzymatic RNA primer extension from the structure of an RNA-GpppG complex

SUPPORTING INFORMATION

Wen Zhang a,b,c,d, Chun Pong Tam a,b,c,e, Travis Walton a,b,c,d, Albert C. Fahrenbach a,b,c,d,f, Gabriel Birrane g, Jack W. Szostak a,b,c,d,e,f,†

 aHoward Hughes Medical Institute, Massachusetts General Hospital, Boston, MA 02114; bDepartment of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114; cCenter for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA 02114; dDepartment of Genetics, Harvard Medical School, Boston, MA 02115; eDepartment of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; fEarth–Life Science Institute, Tokyo Institute of Technology, Tokyo 152-8550, Japan; and gDivision of Experimental Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215

†To whom correspondence should be addressed. Email: szostak@molbio.mgh.harvard.edu.

CONTENTS

1. General Methods
2. Synthesis of Dinucleotides
3. Binding affinity of GpppG with RNA duplex P/T 2c
4. K M measurement for the 2-aminoimidazole bridged intermediate
5. X-Ray Crystallography
6. References
1. General Methods.

1a. General considerations. All reagents for the synthesis of 5'-5'-linked dinucleotides were purchased from Sigma Aldrich (St Louis, MO) and used without further purification, except as noted below. Reactions were conducted in oven-dried 4 mL borosilicate glass vials that were fitted with Teflon-lined caps unless otherwise noted. Reagents and materials used for solid-phase RNA polymerization chemistry, including 50 μmol-scale universal controlled-pore glass (CPG) solid support columns, 5'-DMTr-2'-TBDMS-protected RNA phosphoramidites (bz-A-CE, ac-C-CE, ibu-G-CE, and U-CE), acetonitrile, 0.25 M ethylthio-1H-tetrazole in acetonitrile (activator solution), 0.02 M iodine in THF/H₂O/pyridine (oxidizing solution), 3% trichloroacetic acid in dichloromethane (deblock solution), cap mix A (THF/acetic anhydride/pyridine 8:1:1), and cap mix B (1-methylimidazole/THF/pyridine 8:1:1) were obtained from Bioautomation (Irving, TX). Reagents for RNA column cleavage, protective group removal and purification, including 28% aqueous ammonium hydroxide, 40% aqueous methylamine, anhydrous dimethyl sulfoxide (DMSO), and triethylamine trihydrofluoride (TEA·3HF), 3 M aqueous sodium acetate, 1-butanol and absolute ethanol, were purchased from Sigma Aldrich. Deuterated solvents were purchased from Cambridge Isotope Laboratories (Tewksbury, MA). Preparatory-scale high performance liquid chromatography (HPLC) was carried out on a Varian Prostar 210 HPLC system, equipped with either a preparative-scale Agilent ZORBAX Eclipse-XDB C18 column (21.2x250mm, 7 µm particle size) for reversed-phase chromatography, or with a ThermoFisher-Dionex DNAPac PA100 strong anion exchange column (22x250 mm, 13.5 µm particle size) for anion-exchange chromatography.

1b. NMR data for synthetic dinucleotides. NMR spectra were recorded on a Varian Inova 400 MHz spectrometer (400 MHz for 1H, 100 MHz for 13C, 161 MHz for 31P; Santa Clara, CA). Proton and carbon chemical shifts are reported in parts per million (ppm) values on the δ scale, internally referenced to residual protium in the NMR solvents (Proton NMR: DHO, δ = 4.79 ppm; Carbon NMR: CD₃OD, δ = 49.0 ppm) (1), while proton-decoupled phosphorus chemical shifts were referenced to trimethyl phosphate (D₂O: δ = 3.8 ppm) (2). All NMR spectra were recorded at 25 ºC. Data were reported as
follows: chemical shift, multiplicity (s = singlet, d = doublet, q = quartet, m = multiplet, br = broad), and integration.

1c. RNA oligonucleotides and pGpG synthesis. RNA oligonucleotides and pGpG were synthesized by standard solid-phase phosphoramidite chemistry on a MerMade 6 RNA/DNA oligonucleotide synthesizer (Bioautomation, Irving, TX). Cleavage and elution of the 5′-DMTr-deprotected products from 50 µmol universal CPG-solid support columns were performed by equilibrating and eluting the solid support material a total of 3 times with a 1:1 mixture of ammonium hydroxide and 40% aqueous methylamine (equilibration time: 3 x 10 m; elution volume: 3 x 5 mL for 50 µmol columns). Removal of protecting groups on the nucleobases and phosphates was carried out by heating the basic eluent for 2.5 h at 65 °C; the resultant clear (or off-white) homogeneous mixtures were first concentrated under reduced pressure at 40 °C for 3 h on a Genevac EZ-2 tabletop speedvac system (Genevac, Stone Ridge, NY), then lyophilized to dryness on a VirTis Sentry 2.0 freeze-drier (SP Scientific, Warminster, PA) at <50 mTorr overnight to afford off-white solid residues. The residues were then resuspended in 2.5 mL of DMSO and 2.5 mL of TEA·3HF, and heated for 2.5 h at 65 °C to remove the TBDMS protecting group on the ribose 2′-hydroxyl group. The mixtures were homogeneous and pale to golden yellow in color. After cooling to room temperature (~30 m), 625 µL of 3 M sodium acetate and 15 mL of 1-butanol were added for RNA precipitation. The precipitates were spun down (4000 rpm, 5 m) and supernatants were removed by decanting. The resulting white solids were washed twice with absolute ethanol; the samples were then dried under high vacuum overnight. Purification of the desired products was carried out by preparative-scale HPLC on a Varian Prostar 210 HPLC system equipped with Agilent ZORBAX Eclipse-XDB C18 column using 25 mM triethylammonium bicarbonate in H₂O (TEAB, pH 7.5) with an increasing gradient of 0 % to 15 % acetonitrile over 30 m. Elution of RNA was monitored by UV absorption at 254 and 280 nm. The desired RNA fractions were collected, pooled and lyophilized to afford a fluffy white powder. The resultant white residues were further purified by preparative-scale strong anion-exchange HPLC on a ThermoFisher-Dionex DNAPac PA100 strong anion exchange column with an increasing gradient of 0 to 100 mM aqueous sodium perchlorate solution over 30 m. The desired RNA fractions were collected, pooled and
lyophilized to afford white solid residues. The residues were washed 3 times with acetone, followed by overnight drying under high vacuum, to afford the desired products (in sodium cation form) as fine white grains.

1d. Locked nucleic acid (LNA)-modified RNA oligonucleotides. The LNA-modified RNA oligonucleotide (LNA in bolded letters) used for crystallographic studies (5’-mCmCmCGACUUAAGUCG-3’) was custom-synthesized by Exiqon Inc. (Woburn, MA), with the 5’-dimethoxytrityl (DMT) groups cleaved and samples preliminarily purified by desalting. HPLC purification followed the same procedures as described in section S1c. Concentrations of the aqueous RNA samples were determined by their UV absorption at 260 nm on a Thermo Scientific Nanodrop 2000c Spectrophotometer (Waltham, MA). The theoretical molar extinction coefficients of the RNA strands used herein at 260 nm were provided by Exiqon Inc. (Woburn, MA).

1e. High-resolution mass spectrometry (HRMS) analyses. UHPLC grade (Optima® Grade, Fisher Scientific) reagents and solvents were used to prepare the aqueous buffers and organic solvents for HPLC-TOF-MS analysis, including water, triethylamine, 1,1,1,3,3,3-hexafluoroisopropanol (HFIP), and methanol. High-resolution mass data were obtained for oligonucleotides and synthetic dinucleotides. The analyses were performed on an Agilent 1200 HPLC system coupled to an Agilent 6220 Accurate-Mass time-of-flight mass spectrometer, with a solvent degasser, temperature-controlled auto sampler, column oven, diode-array detector, and a dual electrospray ionization source. Samples were analyzed over a 100 mm XBridge C18 column (1 mm i.d., 3.5 µm particle size, Waters Corporation) using reverse-phase ion-pairing chromatography (3); solvent A was water with 200 mM HFIP, and 1.25 mM TEA at pH 7.0 and solvent B was methanol. For oligonucleotides, solvent B was ramped from 2.5% to 20% over 30 m at a flow rate of 0.1 mL min⁻¹ with the column heated to 50 °C. For dinucleotides, isocratic elution with 2% MeOH in solvent A was used. Typically, 100–200 pmole of the analyte was injected for analysis in extended dynamic range in negative ion mode using the following settings: scan rate, 1 spectrum s⁻¹; mass range, 239 m/z – 3200 m/z; drying gas flow, 8 L m⁻¹; drying gas temperature, 325 °C; nebulizer pressure, 30 psig; capillary voltage, 3500 V; fragmentor, 200 V; and skimmer, 65 V. Data analysis was performed using MassHunter Qualitative Analysis (Agilent Technologies). The UV-Vis absorption
and total ionization count (TIC) traces of the samples were monitored and used to gauge purity; all RNA samples used in this study were determined to have a purity of >90%.

1f. NMR Titration Studies. The initial duplex solution contained 1.5 mM RNA duplex, 500 mM sodium chloride, and 10% D$_2$O; the ligand solution contained the appropriate concentration of GpppG (sodium salt form), as well as the same concentration of the RNA duplex (1.5 mM), D$_2$O (10%), and sodium cation (500 mM) as that of the starting duplex solution, in order to maintain a constant duplex concentration and ionic strength throughout the titration experiment. A monomer solution containing 30 mM of GpppG was assumed to contain 90 mM of Na$^+$, and 410 mM of sodium chloride was added to the monomer solution to bring the concentration of Na$^+$ of the monomer solution to 500 mM. The pH of both duplex and monomer solutions was adjusted to 7.0 ($\pm$ 0.1) using trace amounts of either aqueous sodium hydroxide or hydrochloric acid. NMR spectra were acquired on a Varian INOVA 400 MHz NMR spectrometer equipped with a broadband PFG (z-gradient) probe. Suppression of the bulk water resonance was achieved using a Watergate pulse sequence (4, 5). Each spectrum was recorded after 128–256 scans, with an optimized delay period on pulse sequence (d1) of 1.0 s and a pulse width (pw) of 15 µs. Initial concentrations of the duplex and monomer solutions were determined by their UV absorption at 260 nm on a Thermo Scientific Nanodrop 2000c Spectrophotometer (Waltham, MA). The theoretical molar extinction coefficients at 260 nm of the RNA strands were calculated with Integrated DNA Technologies' OligoAnalyzer 3.1 (Coralville, IA)(6, 7); the molar extinction coefficients of pGpG, GppG, and GpppG at 260 nm were assumed to be 24160 L mol$^{-1}$ cm$^{-1}$. The chemical shifts were referenced externally using a co-axial NMR tube containing a solution of pentafluorobenzaldehyde ($\delta$ = 10.285 ppm) in CDCl$_3$. 

2i. 5′-O-Phosphoguananylyl-(3′→5′)-guanosine (pGpG)

Using standard solid-phase phosphoramidite polymerization chemistry, pGpG was prepared by four 50-μmol-scale solid-phase syntheses (total scale: 0.2 mmol) on a MerMade 6 oligonucleotide synthesizer, in accordance to procedures laid out in section S1c of this Supporting Information.

| 1H NMR (D₂O) | δ 8.11 (s, 1H), 7.99 (s, 1H), 5.84 (d, J = 6.6 Hz, 1H), 5.82 (d, J = 5.6 Hz, 1H), 4.7 (dd, J = 5.7 & 5.7 Hz), 4.46 (dd, J = 3.9 & 4.3 Hz, 1H), 4.41 (br s, 1H), 4.28 (br s, 1H), 4.14 (br s, 2H), 3.92 (br s, 2H). Protons on the 2′-position of the two riboses (~ δ 4.84 – 4.80) were poorly resolved from the HOD signal at δ 4.79 ppm. |
| 13C NMR (D₂O) | δ 162.20, 162.05, 156.89, 156.78, 152.80, 152.65, 138.02, 137.88, 117.48, 117.40, 87.71, 86.93, 84.88 (dd, J = 9 & 3 Hz), 84.65 (d, J = 9 Hz), 76.33 (d, J = 5 Hz), 74.51, 74.03 (d, J = 5 Hz), 71.38, 66.12 (d, J = 5 Hz), 64.5 (d, J = 4 Hz) |
| 31P NMR (D₂O) | δ 0.34, -4.54 |
| HRMS (m/z) | Calc’d for C_{20}H_{25}N_{10}O_{15}P_{2} [M–H]⁻: 707.0982; Found: 707.1016 |

2ii. P₁P²-Diguanosine-5′-pyrophosphate (GppG)

The synthetic protocol was adapted from the report of Tanaka et al. (8). 2-Chloro-1,3-dimethylimidazolinium chloride (DMC, 340 mg, 2 mmol, 2 equiv.; very hygroscopic), imidazole (272.3 mg, 4 mmol, 4 equiv.) and guanosine 5′-monophosphate disodium salt (GMP•2Na⁺, 407.2 mg, 1 mmol, 1 equiv.) were charged into an oven-dried 4 mL borosilicate glass vial equipped with a stir bar and a Teflon-lined cap. D₂O (2 mL, 0.5 M) was added as a solvent. The mixture was stirred at 40 °C for 1 h, followed by addition of another 1 equiv. of GMP•2Na⁺ (407.18 mg). After further stirring at 40 °C overnight, the reaction mixture was diluted 20-fold with water, followed by preparative-scale HPLC purification on a Varian Prostar 210 HPLC system equipped with Agilent ZORBAX
Eclipse-XDB C18 column using 25 mM triethylammonium bicarbonate in H₂O (pH 7.5) with an increasing gradient of 0 % to 15 % acetonitrile over 30 m. Elution of RNA was monitored by UV absorption at 254 and 280 nm. The desired RNA fractions were collected, pooled and lyophilized to afford a fluffy white powder. The resultant white residues were again purified with preparative-scale strong anion-exchange HPLC with a ThermoFisher-Dionex DNAPac PA100 strong anion exchange column with an increasing gradient of 0 to 100 mM aqueous sodium perchlorate over 30 m. The desired fractions were collected, pooled and lyophilized at <50 mTorr to afford white solid residues. The residues were washed with acetone for 3 times, followed by overnight drying under high vacuum, to afford the desired products (in sodium cation form) as fine white grains.

\[ ^1\text{H NMR (D}_2\text{O)} \quad \delta 7.94 (s, 2H), 5.78 (d, J = 5.2 \text{ Hz}, 2H), 4.62 (dd, J = 5.1 & 5.1 \text{ Hz}, 2H), 4.44 (dd, J = 4.4 & 4.6 \text{ Hz}, 2H), 4.31–4.26 (m, 4H), 4.22–4.17 (m, 2H) \]

\[ ^{13}\text{CNMR (D}_2\text{O)} \quad \delta 159.5, 154.6, 152.1, 138.1, 116.9, 88.3, 84.1 (t, J = 4.5 \text{ Hz}), 74.9, 70.9, 65.9 \]

\[ ^{31}\text{P NMR (D}_2\text{O)} \quad \delta -10.50 \]

HRMS (m/z) Calc'd for C₄₀H₂₅N₁₀O₁₅P₂ [M–H]⁻: 707.0982; Found: 707.1013

2iii. \( P^1, P^3\)-Diguanosine-5'-triphosphate (GpppG)

The synthetic procedures leading to GpppG were highly analogous to that of GppG, except that guanosine 5'-diphosphate (sodium salt form) was added in lieu of GMP•2Na⁺ following the initial one-hour DMC-mediated GMP activation.

\[ ^1\text{H NMR (D}_2\text{O)} \quad \delta 8.05 (s, 2H), 5.88 (d, J = 4.8 \text{ Hz}, 2H), 4.67 (dd, J = 4.9 & 4.9 \text{ Hz}, 2H), 4.5 (dd, J = 4.6 & 4.7 \text{ Hz}, 2H), 4.36 (br, 2H), 4.33 (m, 4H) \]

\[ ^{13}\text{CNMR (D}_2\text{O)} \quad \delta 159.4, 154.6, 152.1, 138.1, 116.8, 88.2, 84.0 (d, J = 9 \text{ Hz}), 74.9, 70.7, 65.7 \]

\[ ^{31}\text{P NMR (D}_2\text{O)} \quad \delta -10.45 (d, J = 9.7 \text{ Hz}), -21.76 (t, J = 17.7 \text{ Hz}) \]

HRMS Calc'd for C₂₂H₂₉N₁₀O₁₀P [M–H]⁻: 787.0645; Found: 787.0675
2iv. 1,3-di-(guanosine-5'-phosphoryl)-2-aminimidazolium (Gp-NH₂Im-pG)

The synthesis of the di-guanosine intermediate began by first synthesizing two monomers, guanosine-5'-phosphoryl-(2-aminimidazole) (2-AmImpG) and guanosine-5'-phosphoryl-(1-hydroxy-7-azabenzotriazole) (GMP-OAt). Aqueous mixtures of GMP•2H⁺ (100 mg, 0.275 mmol, 1 equiv.) with either 2-aminimidazole (dark brown oil, 91 mg, 1.1 mmol, 4 equiv.) or 1-hydroxy-7-azabenzotriazole (200 mg, 1.47 mmol, 5.3 equiv.), and triethylamine (200 µL, 1.4 mmol, 4 equiv., ρ = 0.726 g mL⁻¹) were first prepared, vortexed and sonicated until complete homogenization, then flash frozen in liquid nitrogen and lyophilized over 5 days at < 50 mTorr. Once dry, the solids were separately resuspended in DMSO (20 mL, 14 mM) and triethylamine (300 µL, 2.15 mmol, 7.8 equiv., ρ = 0.726 g mL⁻¹). 2,2'‐dipyridyl disulfide (1.2 g, 5.5 mmol, 20 equiv.) and triphenylphosphine (1.2 g, 4.6 mmol, 17 equiv.) were added and the reactions were left stirring overnight (~ 12 hrs). Extra 2,2'-dipyridyl disulfide (0.62 g, 2.8 mmol, 10.2 equiv.) and triphenylphosphine (0.6 g, 2.3 mmol, 8.4 equiv.) were added to the reactions. Four hours later, both reactions were separately precipitated in precipitation solutions containing 120 mL acetone, 60 mL diethyl ether, and 4.5 g of sodium perchlorate. Pellets were washed twice with acetone, followed by house vacuum-drying overnight. Afterwards, one half of the crude monomer GMP-OAt and one half of the crude 2-AmImpG were mixed together in 5 mL H₂O. This mixture was incubated for 70 m at room temperature and became viscous. The mixture was then purified by reverse phase flash chromatography, by direct loading onto a 30 g C18aq column on a Combiflash Rf-200 from Teledyne Isco. The products were eluted over 15 column volumes, with a gradient of 20 mM TEAB (pH = 7.5) over 0 to 20% acetonitrile at a flow rate of 20 mL/m. The fraction containing the di-guanosine intermediate was immediately flash frozen and lyophilized at -20°C on a VirTis AdVantage Plus EL-85 lyophilizer from SP Scientific. Excess triethylamine from the previous purification was removed by reverse phase chromatography again, using the same procedure, except that the aqueous solvent was water instead of TEAB. The fraction containing the intermediate was then flash frozen and lyophilized at -20°C.
1H NMR (D₂O)  δ 7.68 (s, 2H), 6.56 (dd, J = 2.6, 1.8 Hz, 2H), 5.59 (d, J = 5.2 Hz, 2H), 4.51 (t, J = 5.2 Hz, 2H), 4.26 (dd, J = 5.0, 4.5 Hz, 2H), 3.97–3.95 (m, 4H), 3.88-3.83 (m, 2H)

13C NMR (D₂O)  δ 159.4, 154.3, 152.2, 150.7 (weak, t, J = 7 Hz), 138.0, 116.8, 116.4–116.6 (m), 87.9, 83.2 (d, J = 8.3 Hz), 73.9, 70.4, 66.5 (d, J = 6.4 Hz)

31P NMR (D₂O)  δ –9.04

HRMS (m/z)  Calc’d for C23H28N13O14P2 [M–H]⁻: 772.1359; Found: 772.1371

3. Affinity of GpppG for the RNA duplex P/T₂c
3a. Derivation of the NMR binding isotherm.(9)

Scheme S1. The proposed binding scheme between P/T₂c duplex with GpppG.

We have previously reported the use of the P/T₂c duplex (Sequence: P: 5’–CUCAAUG–3’; T₂c: 5’–CCCAUUGAG–3’) to monitor the binding affinity of two consecutive GMP molecules onto the RNA duplex (10). Upon GMP binding, the imino proton signal of G7 (the 3’-terminal nucleotide of the primer, flanking the ligand binding sites) was observed to shift upfield, without observation of new imino proton signals. This observation is in line with our hypothesis that GMP binding with RNA duplexes is in fast exchange, with the on- and off-rates of ligand binding being faster than the NMR time scale. When these conditions are met, and when the ligand-duplex binding stoichiometry is expected to be 1:1, the following single-site binding isotherm is commonly used to numerically approximate the binding affinity (K):

\[ \Delta \delta = \frac{\Delta \delta_{tot} K[M]}{1 + K[M]} \]

with \( \Delta \delta \) being the measured chemical shift change induced by monomer binding, \( \Delta \delta_{tot} \) being the theoretical total chemical shift change when all duplex binding sites are fully
saturated, and [M] being the concentration of free, unbound monomer (not the concentration of monomer added to the duplex solution). Since we expected the binding of GpppG onto P/T\textsubscript{2c} to be tight, we hypothesized that a substantial portion of the added monomer would be bound to the duplex, with only a small fraction of the added monomer being left in the unbound state. We therefore need to derive an equation which will explicitly solve the concentration of unbound monomer at all points of the titration.

We begin the derivation by realizing that the binding of GpppG with the P/T\textsubscript{2c} duplex can be represented by the following equilibrium:

\[ [D] + [M] \rightleftharpoons K [DM] \]

where D represents P/T\textsubscript{2c} duplex, M represents GpppG, and DM represents GpppG-bound RNA duplex. K is equal to:

\[ K = \frac{[DM]}{[D] \cdot [M]} \]

At any point of the titration, the total concentration of duplex (D\textsubscript{tot}) is equal to the sum of the concentration of GpppG-bound RNA duplex (DM), and free RNA duplex (D). The same applies to monomer as well. Hence:

\[ [D] + [DM] = [D\textsubscript{tot}]; \quad [M] + [DM] = [M\textsubscript{tot}] \]

Substituting these equations into the equilibrium expression, we get:

\[ K = \frac{[DM]}{(D\textsubscript{tot} - [DM]) \cdot (M\textsubscript{tot} - [DM])} \]

Rearrangement of this expression in terms of [DM] gives a quadratic equation:

\[ [DM]^2 - \left(M\textsubscript{tot} + D\textsubscript{tot} + \frac{1}{K}\right) [DM] + M\textsubscript{tot}D\textsubscript{tot} = 0 \]

and the roots (solution) of this quadratic expression can be expressed as the following:

\[ [DM] = \left(M\textsubscript{tot} + D\textsubscript{tot} + \frac{1}{K}\right) \pm \left[\left(M\textsubscript{tot} + D\textsubscript{tot} + \frac{1}{K}\right)^2 - 4M\textsubscript{tot}D\textsubscript{tot}\right]^{\frac{1}{2}} \]

As K approaches positive infinity, the two roots bifurcate into M\textsubscript{tot} (larger root) and D\textsubscript{tot} (smaller root); however, since M\textsubscript{tot} > D\textsubscript{tot} at the end of the titration, [DM] must be limited by the total duplex concentration (D\textsubscript{tot}). Hence [DM] cannot be equal to M\textsubscript{tot}, and the larger root is rejected. Hence,
\[ [DM] = \frac{(M_{tot} + D_{tot} + \frac{1}{K}) - \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \]

Subsequently, at any point of the NMR titration, the concentration of free, unbound monomer is equal to

\[ [M] = [M_{tot}] - [DM] \]

\[ [M] = \frac{M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}}}{2} \]

This equation is substituted back into the canonical NMR binding isotherm:

\[ \Delta\delta = \frac{\Delta\delta_{tot}K[M]}{1 + K[M]} \]

and we get the crude NMR binding isotherm which is then used to fit the NMR data for numerical approximation of the GpppG–RNA duplex binding affinity:

\[ \Delta\delta = \frac{\frac{1}{2}\Delta\delta_{tot}K \left\{ M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}} \right\}}{1 + \frac{1}{2}K \left\{ M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}} \right\}} \]

\[ \Delta\delta = \frac{\Delta\delta_{tot}K \left\{ M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}} \right\}}{2 + K \left\{ M_{tot} - D_{tot} - \frac{1}{K} + \left[\left(M_{tot} + D_{tot} + \frac{1}{K}\right)^2 - 4M_{tot}D_{tot}\right]^{\frac{1}{2}} \right\}} \]

The true total monomer concentration \((M_{tot})\) throughout the titration may deviate from the theoretical value, due to factors like handling error in monomer addition, or inaccuracies in the approximation of the true concentration of the ligand in the monomer solution. To account for these possible deviations, the stoichiometric factor \(n\) is introduced to the \(D_{tot}\) term, yielding the final form:
3b. Titration of GpppG into the P/T2c duplex: NMR studies.

Figure S1. Titration of the P/T2c duplex with GpppG while monitoring the 10 to 15 ppm region of the $^1$H NMR spectra. (11, 12) **A.** The component spectra for the GpppG–P/T2c titration were stacked for convenient visualization. The 5′-CC overhang of P/T2c engages in binding with GpppG. All spectra were recorded at 12 °C in the presence of 500 mM Na$^+$ and 10% D$_2$O at pH 7 with a duplex concentration of 1.5 mM. The signal at 10.285 ppm is that of the pentafluorobenzaldehyde reference. **B.** Using the spectral data shown in figure S1A, the change in chemical shifts of all P/T2c imino protons were plotted against the concentration of GpppG. The imino proton signal of G7 (red) shifted the most, followed by the signals of U6 (purple), G1 (blue), and U2 (mustard yellow). The signals of internal imino protons U5, G3, and U4 (crimson, green, and cyan, respectively) shifted only minimally.

To assess the affinity of GpppG with cognate RNA duplexes, we used a previously-reported P/T2c duplex(10) with a 5′-CC overhang. Under the conditions of the NMR titration experiments (500 mM Na$^+$, pH 7, 12 °C), the primer and template strands are known to be stably annealed to form an A-form duplex. Additionally, the imino protons of the P/T2c duplex are clearly defined and baseline-resolved. The assignment

\[
\Delta \delta_{\text{tot}} = \frac{\Delta \delta}{2 + K} \left\{ M_{\text{tot}} - nD_{\text{tot}} - \frac{1}{K} + \left[ \left( M_{\text{tot}} + nD_{\text{tot}} + \frac{1}{K} \right)^2 - 4nM_{\text{tot}}D_{\text{tot}} \right]^{1/2} \right\}
\]
of proton resonances (figure S1A) are deduced from data of previously-described variable-temperature proton and two-dimensional $^1$H–$^1$H NOESY experiments(10, 13).

We carried out a titration from 0 to 9 mM GpppG (0 to 6 equivalents) into 1.5 mM of $P/T_{2c}$ solution, and the change in chemical shift of the G7 imino proton was fitted to a modified single-site binding isotherm. As the concentration of GpppG increased, G7 shifted upfield remarkably (~0.35 ppm, red trace, figure S1B) and in a hyperbolic fashion, which is in line with our previous observations that purine nucleotide ligands tend to shift the G7 resonance upfield (13). The imino proton resonance of U6 also undergoes an upfield and hyperbolic shift, although to a lesser degree (~0.10 ppm, purple trace, figure S1B). On the other hand, both G1 and U2 imino protons (blue and mustard yellow traces, figure S1B) undergo broadening and upfield shifting as GpppG is titrated into the duplex solution, although these changes are much smaller than those seen for the G7 imino proton. Since both G1 and U2 imino protons are on the blunt-end terminus of $P/T_{2c}$ with no GpppG binding sites on the template, we hypothesize that these experimental observations likely arise from side-on, non-specific association of GpppG onto the blunt-end terminus of $P/T_{2c}$, and is likely unrelated to the Watson Crick based GpppG–$P/T_{2c}$ binding of interest. Finally, the three “internal” imino protons (G3, U4, and U5; green, cyan, and crimson traces, figure S1B) only shifted upfield minimally (< 0.05 ppm), showing that the effect of GpppG binding does not propagate beyond two base pairs.

It is interesting to note that, as the concentration of GpppG in the duplex solution is increased beyond 6 mM (asterisked spectrum, figure S1A), all imino proton resonances begin to broaden into the baseline. We are reluctant to speculate on the factors or nonidealities that give rise to these observations; however, we suspect that when GpppG is in excess of available duplex binding sites, and when temperature is lowered, the nonspecific interactions between GpppG and surrounding $P/T_{2c}$ duplexes could be significant enough to be able to bridge multiple RNA duplexes noncovalently, to form a relatively large RNA-GpppG complex that undergoes slow tumbling on the NMR timescale.

S13
4. $K_M$ measurement for the 2-aminoimidazolium-bridged diguanosine dinucleotide (Gp-NH$_2$Im-pG)

4a. Primer extension reaction.

Nonenzymatic template-directed RNA polymerization was monitored through primer extension assays. For accurate comparison between the RNA•GpppG and RNA•Gp-NH$_2$Im-pG, we used the same primer-template duplex as the one used in the NMR binding assay (Sequences: primer, 5’-FAM-CUCAAUG-3’; template, 5’-CCCAUUGAG-3’). The final concentrations of the primer extension reaction were 2 µM primer, 3 µM template, 200 mM Tris pH 8, 100 mM MgCl$_2$, and 0.125–2 mM of the Gp-NH$_2$Im-pG intermediate, adjusted for purity. After adding Gp-NH$_2$Im-pG to the reaction mixture to initiate primer extension, 1 µL reaction aliquots were removed at 1, 2, and 3 m and quenched in 7 µL of 8 M urea, 100 mM Tris-Cl, 100 mM boric acid, and 75 mM EDTA. Reaction aliquots were flash frozen on dry ice, and kept under −80 °C until analysis by 20% denaturing polyacrylamide gel electrophoresis (National Diagnostics, Atlanta, GA). Gels were imaged on an Amersham Typhoon scanner from General Electric and quantified using the ImageQuantTL software (Little Chalfont, United Kingdom). Data were analyzed assuming pseudo-first order kinetics to determine the initial rate constant, $k_{obs}$.

4b. Michaelis-Menten analysis of RNA:Gp-NH$_2$Im-pG binding

To compare the RNA binding affinity of GpppG with that of Gp-NH$_2$Im-pG, we sought to determine the $K_M$ of this intermediate in primer extension reactions. We were able to obtain relatively pure Gp-NH$_2$Im-pG through the synthetic route laid out in Section S2, (page S8) via two rounds of purification, with ≥ 80% of the purified material in the form of imidazolium-bridged dinucleotide, and the remainder as the monomer (2-AmImpG). In addition, these fractions contained 1ed–2 equivalents of triethylamine. For our assays, we assumed that the RNA binding affinity of the intermediate would be greater than that of the monomer, and that the monomer would contribute insignificantly to the polymerization rate under these conditions.

For the primer extension assay, we used the same primer and template RNA sequences as that in the affinity measurement made by NMR. We observed that the
buffer and pH conditions of the NMR studies (pH 7, no Mg$^{2+}$) were not compatible with our $K_M$ determination because of the slow polymerization rate of primer extension. In addition, we observed that solutions containing even 100 mM NaCl resulted in precipitation of the intermediate and poor kinetic analysis of primer extension. Therefore, we used 200 mM Tris pH 8 and 100 mM MgCl$_2$ for our primer extension studies. We determined the initial rate of primer extension for 5 concentrations of the Gp-NH$_2$Im-pG and measured $K_M = 619 \pm 60 \mu$M (figure S2). This value is 3x greater than the $K_d = 175 \mu$M of the GpppG analog by NMR, but is 27x less than the $K_d = 17$ mM of GMP. This suggests that the binding of the intermediate is better approximated by GpppG than GMP. The discrepancy between the $K_M$ of the intermediate and the $K_d$ of GpppG may be due to a variety of factors. For instance, the buffer conditions and pH of the two experimental systems differ. In addition, the relationship between $K_M$ and $K_d$ may not be directly comparable. Therefore, we believe that these values are in rough agreement and support our proposal that GpppG is an analogue of the di-guanosine intermediate.

![Figure S2](image.png)

**Figure S2.** Determination of the $K_M$ of Gp-NH$_2$Im-pG dinucleotide by the primer extension assay. (A) Michaelis-Menten plot of the $k_{obs}$ of primer extension versus the concentration of the di-guanosine intermediate. Black circles represent average experimental values in triplicate and error bars indicate ± 1 S.D. The red line is our fit of the experimental data using $K_M = 0.619$ mM and the maximum $k_{obs} = 9.87$ h$^{-1}$. (B) Lineweaver-Burke plot of the experimental data from part A. Black circles represent experimental data with error bars ± 1 S.D. The red line is the linear regression of this data.
5. X-ray crystallography.

5a. Crystal preparation. The Nuc-Pro High Throughput Screen Kit (Jena Bioscience), Natrix High Throughput Kit and Index High Throughput Kit (Hampton Research, Aliso Viejo, CA) were used for screening crystallization conditions by the sitting drop vapor diffusion method. Solutions containing the RNA sample (0.5 mM) and different dinucleotide ligands (10 mM) were heated to 90 °C for 2 m, then cooled slowly to room temperature. All of the crystals grew at 18 °C, and the mother liquor containing 50% glycerol was used as a cryoprotectant during crystal mounting. All data collection was taken under a stream of nitrogen at 99 K. The data sets were collected at the SIBYLS beamline 821 and 822 at the Advanced Light Source, Lawrence Berkeley National Laboratory. The distances between the detector and the crystal were set to 200 mm and the collecting wavelength was set to 0.997 Å. The crystals were exposed for 1 second per image with one degree oscillations, and 180 images were taken for each data set. The optimized crystallization conditions for the RNA-monomer complexes are listed below.

<table>
<thead>
<tr>
<th>Optimized crystallization conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA–GppG 0.05 M Magnesium chloride, 0.1 M Imidazole pH 6.5, 1.0 M Sodium acetate trihydrate</td>
</tr>
<tr>
<td>RNA–GpG 0.05 M Magnesium chloride, 0.2 M Ammonium citrate tribasic pH 7.0, 20% w/v Polyethylene glycol 3,350</td>
</tr>
<tr>
<td>RNA–pGpG 0.05 M Magnesium chloride, 0.1 M HEPES sodium pH 7.5, 2% v/v Polyethylene glycol 400, 2.0 M Ammonium sulfate</td>
</tr>
<tr>
<td>RNA–GpppG 0.05 M Magnesium chloride, 1.2 M Lithium sulfate, 50 mM MES pH 6.5, 2 mM Cobalt (II) chloride</td>
</tr>
</tbody>
</table>

5b. Data collection and structure refinement. The data were processed using HKL2000 and DENZO/SCALEPACK. All of the structures were solved by molecular replacement, using structure of 5DHC as search model. All four structures were refined using Refmac. The refinement protocol included simulated annealing refinement, restrained B-factor refinement, and bulk solvent correction. During refinement, the
topologies and parameters for locked nucleic acids (LCC) and for the ligands GpppG (GP3), GppG (GP2), GppppG (GP4) were constructed and applied. After several cycles of refinement, a number of highly ordered water molecules and magnesium ions were added. The 4 crystal structures were determined to resolutions of 1.9 Å, 1.5 Å, 2.6 Å and 2.1 Å, respectively (PDB ID: 5UEE, 5UED, 5UEG, 5UEF). Data collection, phasing, and refinement statistics of the determined structures are listed in Tables S2 and S3.

Table S2. Data collection statistics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P3;21</td>
<td>P3;21</td>
<td>P3</td>
<td>P3;21</td>
</tr>
<tr>
<td>Unit cell parameters (Å, °)</td>
<td>46.96, 46.96, 83.10</td>
<td>43.91, 43.91, 85.41</td>
<td>48.42, 48.42, 81.96</td>
<td>43.63, 43.63, 84.05</td>
</tr>
<tr>
<td>Resolution range, Å (last shell)</td>
<td>50-1.90 (1.97-1.90)</td>
<td>50-1.50 (1.55-1.50)</td>
<td>50-2.60 (2.69-2.60)</td>
<td>50-2.10 (2.18-2.10)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>8742</td>
<td>27933</td>
<td>12517</td>
<td>10461</td>
</tr>
<tr>
<td>Completeness, %</td>
<td>99.4 (94.4)</td>
<td>95.1 (100)</td>
<td>96.6 (74.8)</td>
<td>99.8 (99.6)</td>
</tr>
<tr>
<td>Rmerge, %</td>
<td>6.3 (46.4)</td>
<td>7.9 (35.1)</td>
<td>10.0 (55.4)</td>
<td>7.4 (46.9)</td>
</tr>
<tr>
<td>&lt;I/σ(I)&gt;</td>
<td>35.3 (1.9)</td>
<td>28.4 (5.5)</td>
<td>15.1 (1.3)</td>
<td>30.9 (2.8)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>9.1 (4.5)</td>
<td>9.8 (7.6)</td>
<td>5.1 (3.6)</td>
<td>9.6 (7.0)</td>
</tr>
</tbody>
</table>

Table S3. Structure refinement statistics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>5UEE</td>
<td>5UED</td>
<td>5UEG</td>
<td>5UEF</td>
</tr>
<tr>
<td>RNA duplex per asymmetric unit</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Resolution range, Å</td>
<td>83.10-1.90</td>
<td>85.41-1.50</td>
<td>81.96-2.60</td>
<td>84.05-2.10</td>
</tr>
<tr>
<td>Rwork, %</td>
<td>23.20</td>
<td>21.03</td>
<td>20.76</td>
<td>23.97</td>
</tr>
<tr>
<td>Rfree, %</td>
<td>27.47</td>
<td>23.22</td>
<td>30.15</td>
<td>30.32</td>
</tr>
<tr>
<td>Number of reflections</td>
<td>8299</td>
<td>14252</td>
<td>6081</td>
<td>5391</td>
</tr>
<tr>
<td>Bond length R.M.S. (Å)</td>
<td>0.016</td>
<td>0.016</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>Bond angle R.M.S.</td>
<td>1.854</td>
<td>1.843</td>
<td>2.037</td>
<td>1.589</td>
</tr>
<tr>
<td>Average B-factors, (Å²)</td>
<td>32.42</td>
<td>18.04</td>
<td>84.06</td>
<td>46.67</td>
</tr>
</tbody>
</table>
Figure S3. (A) Local structure of RNA-GpppG complex. Two of the water molecules observed to bridge the three neighboring RNA duplexes interact by hydrogen bonding with 2'-hydroxyls of G4s and the G4-G5 phosphodiester phosphates. (B) and (C) Local structures of the RNA-GppG complex. At one end, GppG forms two Watson-Crick base pairs with the template, and at the other end, one Watson-Crick base pair and one noncanonical base pair are observed.

Figure S4. (A) The RNA-GppppG complexes slip-stack to crystallize. (B) The GppppG ligand forms two Watson-Crick base pairs with the template. (C) The RNA-pGpG complexes slip-stack to crystallize. (D) The pGpG dimer forms two Watson-Crick base pairs with the template.
6. References.


