A naturally occurring, noncanonical GTP aptamer made of simple tandem repeats

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.4161/rna.28798

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:17295593

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
A naturally occurring, noncanonical GTP aptamer made of simple tandem repeats

Edward A Curtis and David R Liu
Department of Chemistry and Chemical Biology and Howard Hughes Medical Institute, Harvard University, Cambridge, MA USA

Keywords: naturally occurring aptamer, tandem repeats, in vitro selection, GTP, RNA, microsatellite

Recently, we used in vitro selection to identify a new class of naturally occurring GTP aptamer called the G motif. Here we report the discovery and characterization of a second class of naturally occurring GTP aptamer, the “CA motif.” The primary sequence of this aptamer is unusual in that it consists entirely of tandem repeats of CA rich sequence motifs such as CAACA arranged as three nucleotides. The majority of the aptamers identified had the potential to form G-quadruplex structures, and further analysis confirmed that these aptamers are abundant, especially in the genomes of eukaryotes. Bioinformatic searches indicate that the CA motif is absent from most archaean and bacterial genomes, but occurs in at least 70 percent of approximately 400 eukaryotic genomes examined. These searches also uncovered several phylogenetically conserved examples of the CA motif in rodent (mouse and rat) genomes. Together, these results reveal the existence of a second class of naturally occurring GTP aptamer whose sequence requirements, like that of the G motif, are not consistent with those of a canonical secondary structure. They also indicate a new and unexpected potential biochemical activity of certain naturally occurring tandem repeats.

Introduction

GTP plays a central role in diverse cellular processes, and at least three percent of the proteins in the human proteome bind GTP. The fundamental role played by GTP in cells, as well as the emerging realization that the functions of biological proteins and RNAs may not be as distinct as once thought, led us to hypothesize that RNA molecules that bind GTP might also be widespread in nature. To test this hypothesis, we used in vitro selection to search a pool of phylogenetically diverse genome-derived RNA fragments for new examples of naturally occurring GTP aptamer. Several active variants of the CA motif aptamer lack the ability to form consecutive Watson-Crick base pairs in any register, while others consist of repeats containing only cytidine and adenosine residues, indicating that noncanonical interactions play important roles in its structure. The circular dichroism spectrum of the CA motif aptamer is distinct from that of A-form RNA and other major classes of nucleic acid structures. Bioinformatic searches indicate that the CA motif is absent from most archaean and bacterial genomes, but occurs in at least 70 percent of approximately 400 eukaryotic genomes examined. These searches also uncovered several phylogenetically conserved examples of the CA motif in rodent (mouse and rat) genomes. Together, these results reveal the existence of a second class of naturally occurring GTP aptamer whose sequence requirements, like that of the G motif, are not consistent with those of a canonical secondary structure. They also indicate a new and unexpected potential biochemical activity of certain naturally occurring tandem repeats.

Results

GTP aptamers made of tandem repeats

Previously, we used in vitro selection to isolate GTP aptamers from a pool of phlogenetically diverse genome-derived RNA fragments (Fig. 1A). The majority of these aptamers had the potential to form G-quadruplex structures, and further analysis

Correspondence to Edward A Curtis; Email: curtis@uochb.cas.cz; and David R Liu; Email: drliu@fas.harvard.edu

http://dx.doi.org/10.4161/rna.28798

Volume 11 Issue 6

RNA Biology 11:6, 682–692; June 2014; © 2014 Landes Bioscience
demonstrated that both RNA and DNA G-quadruplexes possess an intrinsic GTP-binding activity. Most of the remaining sequences isolated in the selection contained short CA-rich motifs such as CAACA that often occurred in the context of tandem repeats (Fig. 1B). To determine whether members of this second family of sequences were also GTP aptamers, three examples were synthesized and tested for the ability to bind GTP. Each example bound GTP-agarose at least 20-fold better than α
random sequence control RNA, and none bound agarose lacking GTP at significant levels (Fig. 1C).

The most effective GTP binder among these aptamers, GG(GCAACA)_6, was chosen for further characterization. This aptamer could be eluted from GTP-agarose columns with GTP, indicating that it binds free as well as immobilized GTP (Fig. 2A). Based on the concentration dependence of elution,12,13 we estimate that this aptamer binds free GTP with a dissociation constant (K_d) of 1.8 mM (Fig. 2B). Elution profiles were consistent with a single binding site for GTP (as was assumed in Fig. 2B), but could be described equally well by more complex models in which multiple, independent binding sites are considered. The aptamer could also be efficiently eluted from GTP-agarose columns with ddGTP, and somewhat efficiently eluted with GDP, but not with 7-methyl-GTP, 6-methylthio-GTP, ITP, ATP, CTP, or UTP (Fig. 2C). These observations indicate that the CA motif recognizes GTP primarily through interactions with its nucleobase (Fig. 2D). Taken together, these results demonstrate that naturally occurring simple tandem repeats can form binding sites specific for biological small-molecule ligands such as GTP.

Unusual sequence requirements of the CA motif aptamer

To better understand the types of tandem repeats that can bind GTP, we used site-directed mutagenesis to characterize the sequence requirements of this aptamer in greater detail. We focused our efforts on the (GCAACA)_n aptamer, both because its sequence was the simplest among the tandem repeat aptamers identified and because it bound GTP the most efficiently (Fig. 1C). To determine the minimum number of repeats needed for efficient binding as well as the number required for maximal binding, a series of aptamer variants containing between two and 16 repeats were synthesized and tested for the ability to bind GTP-agarose. No significant binding was observed for variants containing two or four repeats (Fig. 3A). The ability to bind GTP increased dramatically for a variant containing six repeats, and continued to increase as repeat number increased to ten, before gradually decreasing (Fig. 3A). We hypothesize that at higher repeat numbers the expected increase in GTP-binding activity from additional binding sites is offset by less efficient aptamer folding.14

To identify nucleotides in the aptamer most important for its GTP-binding activity, variants containing each of the 18

![Figure 2. Characterization of the affinity and specificity of the CA motif aptamer using competitive column elution. (A) Elution of the CA motif aptamer from a GTP-agarose column with free GTP. (B) Elution of the CA motif aptamer from a GTP-agarose column as a function of free GTP concentration. (C) Elution of the CA motif aptamer from a GTP-agarose column in the presence of various GTP analogs. (D) Competitive column elution data mapped onto the chemical structure of GTP. The CA motif construct GG(GCAACA)_6 was used for these experiments. Values reflect the average of three independent experiments, and error bars indicate one standard deviation.](image-url)
possible single-mutation changes in the GCAACA motif were synthesized and tested for the ability to bind GTP in the context of an aptamer containing six tandem repeats. These results confirmed the importance of the CA-rich portion of the sequence. While in some cases mutational changes at the first, second, or third positions had only small effects, all single-mutation changes at the fourth, fifth, and sixth positions either strongly decreased or abolished the GTP-binding activity of the aptamer (Fig. 3B). This analysis also revealed that, at least in the context of single-mutation changes, the optimal nucleotide at five out of six positions in the GCAACA repeat is either a cytidine or adenosine (Fig. 3B).

To further explore this hypothesis, we investigated the extent to which the CA motif of the GCAACA repeat only requires cytidine and adenosine. Taken together, these results suggest that the CA motif of the GCAACA repeat only requires cytidine and adenosine. To further explore this hypothesis, we investigated the extent to which the CA motif of the GCAACA repeat only requires cytidine and adenosine.

CA motif aptamer repeats only require cytidine and adenosine.

Taken together, these results suggest that the CA motif of the GCAACA repeat binding to its aptamer may not require canonical base pairs for its activity. To further explore this hypothesis, we investigated the extent to which the CA motif of the GCAACA repeat only requires cytidine and adenosine.

Figure 3. Sequence requirements of the CA motif aptamer. (A) Binding of the GG(GCAACA)₆ aptamer to GTP-agarose as a function of repeat number. (B) Ability of all single-mutation variants of the GCAACA motif in the GG(GCAACA)₆ aptamer to bind GTP-agarose. (C) Ability of all single-mutation variants of the GCAACA motif in the GG(GCAACA)₆ aptamer to bind GTP-agarose. Percent bound values reflect the average of three independent experiments, and error bars indicate one standard deviation.
To more clearly elucidate the nature of these interactions, we used site-directed mutagenesis to search for compensatory mutational effects among positions 2, 3, 4, and 6 in the GG(GCAACA)_6 aptamer. Fifteen of the 16 possible variants differed significantly from that of tRNA (A-form RNA), although we note that considerable heterogeneity in such spectra has been reported. Like the CD spectrum of the CA motif, the spectra of these triplets sequences contain negative peaks at ~210 nm and ~250 nm, a positive peak at ~270 nm.21-24 Taken together, these results provide additional evidence that the CA motif does not form a canonical A-form duplex, and are consistent with an unusual structure such as a triple-stranded helix.

Phylogenetic distribution and evolutionary conservation of the CA motif aptamer

Our original selection revealed that the CA motif occurs in both the human and chicken genomes, but provided little information about other species in which this aptamer is found. To better characterize the phylogenetic distribution of the CA motif, we searched genomic sequence databases for additional examples using a sequence model derived from the site-directed mutagenesis experiments described in Figures 3A and B, 4B, and 5A. This model required sequences to contain at least six copies of the GCAACA motif or at least ten copies of the ACAACA motif, allowed for five mutational changes (1G to U, 2C to A, 2C to D, 3A to C, and 1G5C to 1C5G) observed in active variants of the GCAACA motif, and required that tandem repeats be comparable to that of the GG(GCAACA)_6 aptamer.

Different classes of nucleic acid structures often exhibit characteristic circular dichroism (CD) spectra. To investigate the relationship between the structure of the CA motif and those of other types of nucleic acid motifs, we obtained the CD spectrum of the GG(GCAACA)_6 reference sequence and four variants of this aptamer. These variants differed from the reference sequence at six positions, and each bound GTP-agarose with an efficiency comparable to that of the GG(GCAACA)_6 aptamer (Fig. 3B). The CD spectra of these five sequences were similar: each contained a negative peak at ~220 nm and ~280 nm, a positive peak at ~270 nm, and a peak that was usually positive at ~220 nm (Fig. 6A). In the presence of 1 mM GTP, spectra were similar to those measured in the absence of GTP, although some differences were observed for the 1U construct (Fig. 6A). In contrast, spectra of CA motif variants differed significantly from that of tRNA (A-form RNA), which contains a negative peak at ~220 nm and a positive peak at ~260 nm (Fig. 6B). They also differed from typical spectra of several other important types of nucleic acid structures, including the B-form helix, the Z-form helix, the parallel strand G-quadruplex, the antiparallel strand G-quadruplex, and the i-motif.28

Circular dichroism spectrum of the GCAACA aptamer

The spectrum of the CA motif most closely resembles those of several sequences that adopt triple-stranded architectures,24 although we note that considerable heterogeneity in such spectra has been reported.22 Like the CD spectrum of the CA motif, the spectra of these triplets sequences contain negative peaks at ~210 nm and ~250 nm and positive peaks at ~220 nm and ~270 nm.21-24 Taken together, these results provide additional evidence that the CA motif does not form a canonical A-form duplex, and are consistent with an unusual structure such as a triple-stranded helix.
contained 35 or more examples. Examination of the species in which the CA motif was most abundant (Table S2) revealed several taxonomic groups in which this aptamer occurred frequently in all available genomes, including family Muridae (Mus musculus and Rattus norvegicus) and family Culicoidea (Anopheles gambiae, Aedes aegypti, and Culex quinquefasciatus). Analysis of gene expression databases such as the Gene Expression Omnibus indicated that at least some of these examples are also expressed. For example, of the 29 species in which the CA motif occurs 25 or more times, 19 contain at least one expressed example supported by EST evidence, and six more contain at least one example predicted to be expressed (Table S2). Because we have likely not identified all possible variants of the CA motif that can bind GTP, the results of these searches represent lower limits on the true number of CA motif aptamers in the genomes analyzed.

In addition to examining the overall abundance of the CA motif in different species, we also investigated the extent to which specific examples have been conserved in evolution. Because the CA motif is abundant in both mouse and rat, and the comparative genomics of this pair is well studied, we focused most of our efforts to identify conserved variants on these two species. Rat orthologs of mouse CA motif variants were identified using whole-genome alignments, manually examined to determine their similarity to the CA motif consensus sequence, and in some cases tested for the ability to bind GTP. Although conserved orthologs were not detected for most of the approximately 120 CA motif variants analyzed, two mouse/rat pairs were identified in which the CA motif was potentially conserved (Fig. 7B). For both pairs, the mouse sequence was an exact match to the CA motif consensus, while the rat ortholog was similar but not identical (Fig. 7B). One of these examples (mouse/rat 1) mapped to an intergenic region on chromosome 5, while the other (mouse/rat 2) occurred antisense to an intron in the Hid1 gene. Both mouse sequences as well as their rat orthologs bound GTP-agarose (Fig. 7B), indicating that in some cases the ability of the CA motif to bind GTP has been conserved in evolution.

Figure 5. CA motif variants with repeats that contain only cytidine and adenosine. (A) Binding of the GG(ACAACA)n aptamer to GTP-agarose as a function of repeat number. (B) Binding of GG(ACAACA)n to control agarose (Con) and GTP-agarose (GTP). Binding of a random sequence control with the sequence GG(N)48 is also shown as the leftmost set of bars. (C) Compensatory CA to AC mutations at positions 2 and 6 in each repeat of the aptamer. Mutations were tested in the context of the 3C background, which has the sequence GG(GCAACA)n. (D) Compensatory CA to AC mutations at positions 2 and 6 in each repeat of the aptamer. Mutations were tested in the context of the reference background, which has the sequence GG(GCAACA)n. (E) Parallel-strand model with the 1–5 and 2–6 constraints indicated by dashes. For simplicity only two copies of the GCCAACA motif are shown. For A–D, percent bound values reflect the average of three independent experiments, and error bars indicate one standard deviation.
Figure 6. For figure legend, see page 689.
Although antiparallel, double-stranded, right-handed helices formed by canonical Watson-Crick base pairs are the most common structural elements in both RNA and DNA, numerous variations on this theme have been identified. These variations include noncanonical base pairs occurring in the context of otherwise standard helical elements, unusual helical architectures such as Z-DNA and triplex DNA, and structures in which canonical base pairs are entirely absent. Examples include the G-quadruplex, which contains guanosine tetrads rather than standard base pairs, and the i-motif, which contains hemiprotonated C-C base pairs organized into a four-stranded structure.

The sequence requirements of the CA motif suggest that, like the G-quadruplex and the i-motif, the structure of this aptamer does not require canonical base pairs. Although our experiments do not provide direct information about the three-dimensional fold of this structure, they significantly constrain the possibilities. Variants of the CA motif made up of repeats containing only cytidine and adenosine can still bind GTP (Fig. 5A and B).
suggesting that the essential interactions made by the aptamer only require these two nucleotides. Compensatory mutations involving CA to AC changes (Figs. 5C and 5D) suggest that in some cases these interactions occur between cytidine and adenosine. The CD spectrum of the CA motif suggests that it does not form a standard A-form helix (Fig. 6B). Finally, if the 1–5 (Fig. 4B) and 2–6 interactions (Fig. 5C and D) reflect physical contacts, they are consistent with a structure containing parallel strands (Fig. 5E).

The structure of the CA motif may be related to that of a viral translational enhancer element called the omega sequence. This enhancer consists almost entirely of imperfect CAA repeats, although the number of repeats is not high enough to be expected to bind GTP under the conditions used in our study. The omega sequence by ultracentrifugation, thermal melting, and chemical probing indicates that, despite the inability to form canonical Watson-Crick base pairs, it adopts a compact and stable structure. The CA-rich region of this enhancer was proposed to form a triple helix containing both parallel and antiparallel strand topologies; and some details of this model are consistent with our data. For example, our site-directed mutagenesis experiments suggest that the CA motif aptamer could form a structure containing parallel strands (Figs. 4B and 5C–E). The CD spectrum of the CA motif (Fig. 6) is also consistent with a triple helix, although a high-resolution structure will be needed to confirm this hypothesis. A second known example of a stable structure formed by an RNA containing only cytidine and adenosine is a dimer formed by two molecules of CA and an intercalated proflavine molecule. Like the proposed fold of the omega sequence, the structure of this dimer contains parallel strands.

The results of our selection using a pool of genome-derived RNA fragments differ significantly from those of most previous studies in which functional RNAs were isolated from random sequence pools. The majority of the aptamers and ribosymes isolated in these studies form secondary structures containing standard helical elements. For example, 14 GTP aptamers previously identified using in vitro selection, including at least six distinct motifs isolated from random sequence pools, form secondary structures containing canonical Watson-Crick base pairs. In contrast, neither of the GTP-binding motifs we isolated from a pool of genome-derived RNA fragments adopts such a structure. This observation is consistent with the idea that nucleic acid motifs that form noncanonical structures may be more abundant in eukaryotic genomes than they are in random sequence pools. It also suggests that efforts to identify naturally occurring functional RNA elements should include methods capable of identifying noncanonical motifs, such as the selection-based approach used here, in addition to those designed to identify phylogenetically conserved RNA secondary structures containing Watson-Crick base pairs. Recent studies suggest that such motifs can bind a wide range of biological cofactors, providing further evidence of the functional diversity of noncanonical nucleic acid structures.

RNA aptamers that bind small molecules are widespread in bacteria, and play important roles in the regulation of gene expression. The results described in this and a previous study are consistent with the possibility that such aptamers also occur more frequently in eukaryotes than is currently appreciated. We anticipate that direct selection-based approaches will continue to serve as a powerful approach to further explore this possibility.

Materials and Methods

In vitro selection

Pools were generated by fragmentation of genomic DNA using DNeasy I followed by gel purification of ~108–600 bp fragments on agarose gels. Fragments with 3′ adenosine overhangs were generated by incubating first with DNA polymerase I, and then with dATP and T7 DNA polymerase. These fragments were ligated into pGEM-T vectors and amplified by PCR using primers flanking the insertion site, one of which contained a T7 promoter at its 5′ end. Templates were transcribed using T7 RNA polymerase to generate starting pools for in vitro selection experiments.

GTP aptamers were isolated by incubating pool RNA with GTP-agarose, washing away unbound molecules with selection buffer, and eluting bound RNAs with EDTA. Eluted molecules were subjected to RT-PCR and transcribed to generate RNA for the next round of selection. After four rounds of selection, the pool was cloned using the TOPO TA kit (Invitrogen) and sequenced. See our previous study for detailed descriptions of the protocols.

GTP-agarose binding assays

Assays were performed as previously described using GTP-agarose from Innova Biosciences and micro-spin columns from Pierce (catalog number 89879). Following the addition of RNA (100 nM final concentrations with a trace amount of body labeled RNA in a volume of 170 µL), columns were incubated for 15 min and then washed four times with 170 µL 1 × binding buffer (20 mM MgCl₂, 200 mM KCl, 20 mM HEPES pH 7.1). Bound RNA was eluted by washing twice with 170 µL of 5 mM EDTA, pH 9.

To determine the percent of RNA bound to GTP-agarose, equivalent percentages of wash and elution fractions were ethanol precipitated, resuspended, and spotted on TLC plates. Plates were scanned using a Typhoon phosphorimager, and counts determined using ImageQuant software. For constructs for which the fraction bound to control agarose was also determined, this value was never significantly different from the fraction bound by a random sequence control pool (Fig. 3C).

Competitive columns elution

The GG(GCAACA)₆ construct of the CA motif aptamer was bound to GTP-agarose using the protocol described above. After washing twice with 170 µL of 1 × binding buffer to remove non-specifically bound RNA, 50 µL of elution buffer (20 mM MgCl₂, 200 mM KCl, 20 mM HEPES pH 7.1 and 5 mM GTP analog) was added to the column. After incubating for 5 min, the column was briefly spun in a benchtop centrifuge, and the flowthrough was transferred to a clean tube. After performing this elution protocol zero (Fig. 2C) or four (Fig. 2A) more times,
the remaining RNA was removed from the column by washing twice with 170 µL 5 mM EDTA. Fractions were precipitated and analyzed as described above. The total number of bound counts was defined as the sum of the counts in all elution fractions plus the sum of the counts in all EDTA washes.

A modified elution buffer (20 mM MgCl₂, 200 mM KCl, 20 mM HEPS pH 7.1, 50 mM NaCl and varying concentrations of GTP) was used for the experiments described in Figure 2B. Elution volumes were either 50 µL (10× in 3 mM GTP) or 20 µL (5× in 10 mM GTP). Results for each GTP concentration. The solubility of GTP in the elution buffer prevented us from using concentrations greater than 10 mM. Elution profiles were fit to the equation

\[ V_e = \frac{(V_n \times F) + (L \times V_f)}{V_n + F(L + K_d)} \]

where \( V_e \) is the median elution volume at the indicated free GTP concentration, \( K_d \) is the dissociation constant for free GTP, \( V_n \) is the elution volume in the absence of free GTP in the elution buffer, \( L \) is the concentration of free GTP in the elution buffer, and \( V_f \) is the median elution volume in the presence of control agarose lacking immobilized GTP. These volumes were then plotted as a function of GTP concentration and fit to the equation

\[ V_e = \frac{(V_n \times F) + (L \times V_f)}{V_n + F(L + K_d)} \]

where \( V_e \) is the median elution volume at the indicated free GTP concentration, \( K_d \) is the dissociation constant for free GTP, \( V_n \) is the median elution volume in the absence of free GTP in the elution buffer, \( L \) is the concentration of free GTP in the elution buffer, and \( V_f \) is the median elution volume in the presence of control agarose lacking immobilized GTP.

Circular dichroism

In a typical experiment, a 200 µM nucleic acid solution in a volume of 100 µL was heated at 65 °C for 5 min and cooled at room temperature for 5 min. After addition of 100 µL of buffer prevented us from using concentrations greater than 10 mM. As described above, the best fit was determined using a value of \( V_e \) that never exceeded 32 µL (the median elution volume at 10 mM GTP, the highest GTP concentration at which the rate of elution was determined).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Mike Lawrence as well as John Guilinger and other members of the group for helpful discussions. This work was supported by HHMI and the NIH/NIGMS (R01 GM069865).

Supplemental Material

Supplemental material may be found here: www.landesbioscience.com/journals/rnabiology/article/28798/


34. Rusten AA, Shirshov NE, Gaddis AT, Spins AS. The leader sequence of tobacco mosaic virus RNA devoid of Wobble-Crick secondary structure possesses a cooperatively modeled complex of formation of the omega RNA. Biochemistry (Mosc) 2013; 78:485-95; PMID:23949479; http://dx.doi.org/10.1021/bi301891p

35. Agalarov SC, Seguin EA, Shirshov NE, Spins AS. Inside the structural organization of the omega leader of TMT RNA. Biochemistry (Mosc) 2007; 76:3005-14; PMID:17979450; http://dx.doi.org/10.1021/bi070846s


