Please cite this article as:

hypothesis

Xinmiao Fu. Potential protein-encoded synthesis of DNA and RNA. *Hypothesis* 2014, **12**(1): e6, doi:10.5779/hypothesis.v12i1.366

State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, China

*Correspondence:
fuxinmiao@pku.edu.cn

Received: 2014/02/27; Accepted: 2014/05/12;
Posted online: 2014/09/12

Potential protein-encoded synthesis of DNA and RNA

Xinmiao Fu*

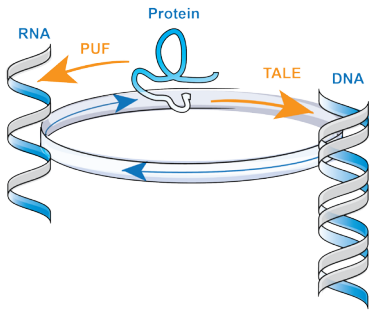

Illustration by Eloïse Kremer

ABSTRACT As stated by the central dogma of molecular biology, in living organisms the sequence information of DNA is transferred to RNA and then to protein, but such information cannot be transferred back from protein to nucleic acids. In this article, it is proposed that the sequence information encoded by protein can be artificially transferred back to DNA and RNA, respectively, based on transcription activator-like effectors (TALE) and Pumilio/fem-3 mRNA-binding factors (PUF). Specifically, mono- and/or dinucleotides are assumed to be arranged along the characteristic amino acids of TALE and PUF, and then assembled as oligonucleotides by ligase or condensation agents. This hypothesis suggests a new protein-based strategy for synthesizing DNA and RNA molecules.

INTRODUCTION The central dogma of molecular biology, as proposed by Francis Crick in 1958[1] and later formulated in 1970[2], is based on extensive genetic and biochemical studies on living organisms and is fundamental to life sciences. It follows that the sequence information of DNA can be self-replicated or transferred residue by residue to RNA and then to protein (**Figure 1A**). However, the sequence information encoded by protein cannot be transferred back to nucleic acids. In some special cases, DNA can also serve as a template in directing in vitro protein synthesis in the presence of antibiotics such as neomycin[3,4]. In addition, regarding the origin of the central dogma, it is widely believed that the codon-amino acid stereochemical pairing occurred during early evolution and played a crucial role in the current ribosome-mediated translation process, which does not involve the direct interaction between codons and corresponding amino acids[5,6,7,8].

In this article, I suggest the possibility that the sequence information encoded by protein can be artificially transferred residue by residue directly to DNA and RNA, respectively, based on transcription activator-like effectors (TALE) and Pumilio/fem-3 mRNA-binding factors (PUF). This hypothesis, if proved experimentally, suggests a new strategy for the synthesis of short DNA/RNA molecules.

HYPOTHESIS *Synthesis of DNA encoded by TALE proteins* TALE proteins are secreted by bacterial plant pathogens and specifically recognize host DNA sequences via their modular DNA-binding domain of tandem repeats[9]. Each repeat comprises around 34 conserved amino acids and targets a specific base pair by using repeat variable diresidues (RVDs) at positions 12 and 13[10,11] (**Figure 1B**). Structural studies have revealed that multiple TALE repeats form a superhelical structure to track along the sense strand of the DNA duplex within its internal layer and that the 12th residue in each repeat stabilizes the RVD loop while the 13th residue makes a nucleotide-specific contact[12]. Since the DNA specificity can be designed by customizable assembly of TALE repeats, TALEN (a TALE protein fused with a nuclease protein) is being deployed as a powerful DNA targeting tool utilized in different organisms[9].

On the other hand, the code of DNA recognition by RVDs of TALE repeats also suggests the possibility that the sequence information of the RVDs in a TALE protein can be transferred residue by residue to DNA. Although the minimal number of repeats in an artificial TALE to significantly activate gene expression in cells is reported to be 6.5[10], it is conceivable that mono- and dinucleotide duplexes can be arranged specifically along the RVD loops of TALE repeats (**Figure 1B**) based on the following reported observations. First, the minimal number of repeats found in naturally occurring TALE proteins is as low as 1.5[13], indicating that a mononucleotide base pair alone is likely to be sufficient to interact with a TALE repeat. Second, each base pair from the DNA duplex[12] was reported to

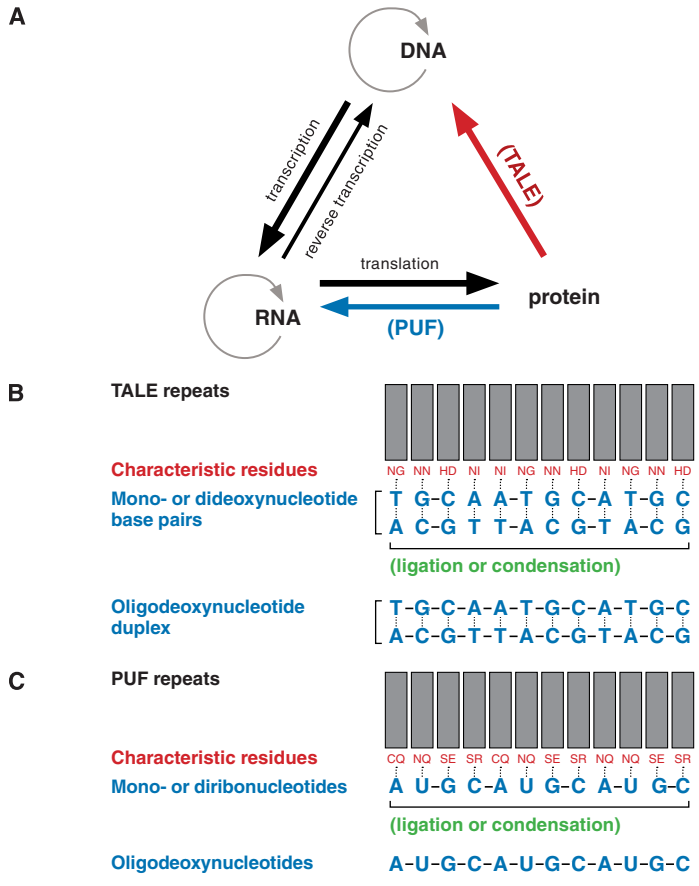

Figure 1 | Illustration of the modified central dogma from potential synthesis of DNA and RNA encoded by protein. A Sequence information transfer (from DNA to RNA to protein) takes place in nature and is summarized by the central dogma of molecular biology (black). Sequence information transfer from protein (e.g., TALE or PUF) to DNA or RNA is illustrated (highlighted in red or blue). **B** Schematic illustration of the synthesis of a DNA duplex by DNA ligase or condensation agents from mono- or dideoxyribonucleotide base pairs that are arranged in a sequence-specific manner along the characteristic residues of TALE repeats in a designed TALE protein. The amino acid code of DNA sequence specificity was adopted from Bogdanove et al.[9]. **C** Schematic illustration of the synthesis of RNA by RNA ligase or condensation agents from mono- or diribonucleotides that are arranged in a sequence-specific manner along the characteristic residues of multiple PUF repeats. The amino acid code of RNA base specificity was adopted from earlier studies[20,21].

independently interact with the RVD loop of one TALE repeat, also strengthening the possibility of an interaction between a TALE repeat and the mononucleotide base pair. Third, mono- or dinucleotide base pairs, although unable to exist stably in solution[14], were reported to be stably formed in the hydrophobic cavities of self-assembled cages[15,16]. Fourth, the apparent dissociation constant between optimized TALE proteins and target DNA was reported to be as low as 0.16 nM[17]. Lastly, TALE repeats were reported to be flexible in conformation to cope with the B-form conformation of the DNA duplex[12].

Together, these observations strongly suggest that mono- or dinucleotide base pairs can be arranged along the corresponding characteristic amino acids of TALE proteins through sequence-specific interactions as described above. As such, these independent but arranged base pairs would be easily ligated or condensed as an oligodeoxynucleotide duplex (as illustrated in **Figure 1B**). In this regard, the sequence information of multiple RVDs discontinuously encoded in the TALE protein is transferred to DNA.

Synthesis of RNA encoded by PUF proteins Similarly, the transfer of sequence information from protein to RNA can potentially be achieved using PUF, which contains an RNA-binding domain that comprises multiple PUF repeats and also forms a superhelical structure to bind the RNA molecule within its inner concave surface[18]. Three-dimensional structural studies[19] revealed that each PUF repeat, though not making direct interactions with either the phosphate backbone or the 2' hydroxyl groups of RNA, recognizes a single RNA base through its three conserved amino acids, two of which make hydrogen bonds or Van der Waals interactions with the edge of an RNA base and a third residue that stacks with the same base and/or the preceding base. The code of RNA-binding specificity by PUF repeats has been decoded and artificially evolved[20,21] (also illustrated in **Figure 1C**), by which modular PUF repeats capable of selectively binding specific RNA sequences can be created.

The bioinformatics analysis results described here indicate that, although a majority contain 8 PUF repeats, many of the naturally occurring PUF proteins only carry 2 or 3 PUF repeats (**Table 1**). In particular, the PUF repeats in some PUF proteins are not sequentially connected but discontinuously present as discrete units of one, two or three repeats (as exemplified in **Figure 1**). These observations indicate that the minimal number of PUF repeats for their stable interaction

Repeat No.	2	3	4	5	6	7	8	9	Total
Protein No.[a]	7	38	94	91	167	91	532	4	1024
Ratio (%)	0.7	3.7	9.2	8.9	16.3	8.9	52	0.4	100

Table 1 | Varying number of PUF repeats in naturally occurring PUF proteins.

[a] A total of 1024 PUF proteins were found in SMART database (http://smart.embl.de/; date of 2013/6/5). The number of PUF repeats in each protein was counted and a summary of PUF proteins with different repeats is presented here.

PUF = Pumilio/fem-3 mRNA-binding factors; SMART = Simple Modular Architecture Research Tool.

with RNA bases may be as low as 2, or even 1. It follows that the mono- and/or diribonucleotides arranged along the concave surface of a designed PUF protein could be assembled as an oligoribonucleotide (as illustrated in **Figure 1C**). As such, the oligoribonucleotide sequence is solely determined by the characteristic amino acids of PUF repeats in the PUF protein.

EXPERIMENTAL DESIGN ***TALE-encoded synthesis of DNA*** To synthesize specific DNA duplexes, the gene encoding multiple TALE repeats plus the N- and C-terminal signals of TALE can be designed and assembled as described previously (as reviewed by Bogdanove et al.[9]) and the protein can be expressed and purified as described recently[12,22]. Mixed 5'-phosphate mono- (4 types) and dideoxyribonucleotides (16 types) are then incubated with the purified TALE protein for a certain length of time. In particular, the binding affinity of the TALE protein to each of these deoxynucleotides can be evaluated by commercial systems based on isothermal titration calorimetry or surface plasma resonance. According to the early reports[12,22] on the interaction between TALE and short DNA duplexes, the preliminary conditions used here are as follows: TALE at a concentration of approximately 2 μM and the mono- and dideoxyribonucleotides at a concentration of approximately 1 mM.

One approach to assembling these mono- and dideoxyribonucleotide base pairs is to use DNA ligase plus ATP and optimal ligation buffer. However, given that the DNA duplex has been reported to be absorbed in the inner surface of the superhelical structure of the TALE protein[12], DNA ligase bearing a molecular weight of 43 kDa may not be able to access those partially buried mono- and dideoxyribonucleotide base pairs due to space hindrance, and thus would not be suitable to catalyze the ligation. An alternative approach to overcome this difficulty is to use chemical agents, such as cyanogen bromide[23], cyanamide[24] and imidazole[25], which have been reported to be able to condense mono- and dideoxyribonucleotide base pairs into oligonucleotide duplexes. Since the base pairs are aligned along the characteristic RVDs of the TALE protein, condensation by these agents should be carried out efficiently due to the entropy reduction.

PUF-encoded synthesis of RNA To synthesize specific single-stranded RNA, the gene encoding multiple PUF repeats can be designed and assembled according to the principles and methods described previously[20,21], and the PUF protein can be expressed and purified similarly as reported previously[19,21]. The binding affinity of the PUF protein to various types of mono- and diribonucleotides can be evaluated by commercial systems based on isothermal titration calorimetry or those based on surface plasma resonance. The results of this kind of experiment will determine the approach and conditions that are suitable for the subsequent ligation and condensation. The preliminary conditions used here are as follows: PUF at a concentration of approximately 2 μM and the mono- and diribonucleotides at a concentration of approximately 1 mM.

If polynucleotide phosphorylase is used for ligation[26,27], a specific diribonucleoside (with a free 3'-terminal hydroxyl group) corresponding to the last two PUF repeats of the PUF protein and mixed nucleoside 3',5'-diphosphates can be added and incubated with the PUF protein. Polynucleotide phosphorylase can be added to the mixture to initiate the addition of mononucleoside 3',5'-diphosphates to the diribonucleoside, which will be further elongated, base by base, to oligoribonucleotides by the enzyme.

A substitute for polynucleotide phosphorylase is RNA ligase[28,29,30], which is known to catalyze the formation of an internucleotide phosphodiester bond between an oligonucleotide acceptor with a 3'-terminal hydroxyl (minimal acceptor being trinucleotides) and an oligonucleotide donor molecule with a 5'-terminal phosphate (minimal donor being trinucleoside diphosphate, dinucleoside pyrophosphate and mononucleoside 3',5'-biphosphates). For this purpose, a specific triribonucleoside (with a free 3'-terminal hydroxyl group) corresponding to the last three PUF repeats of the PUF protein serves as the starting primer, and dinucleoside pyrophosphate or mononucleoside 3',5'-biphosphates serve as building blocks. Specifically, if the building block is mononucleoside 3',5'-biphosphate, ATP should be included for ligation; if it is dinucleoside pyrophosphate, ATP is not required.

The third approach to assemble these mono- and diribonucleotides into oligoribonucleotides is to use condensation agents such as cyanogen bromide[31] and montmorillonite[32,33]. To this end, mixed 5'-biphosphate mono- and diribonucleosides are incubated with the PUF protein for a certain length of time before adding these condensation agents and other crucial chemicals (e.g., metal ions).

SIGNIFICANCE AND CONCLUSION In summary, it is proposed that sequence-specific DNA and RNA molecules can possibly be synthesized according to the characteristic amino acid sequences of designed TALE and PUF proteins, respectively. In terms of thermodynamics, TALE and PUF proteins arrange or fix the free mono- or dinucleotides, leading to a reduction in the entropy of the nucleotides and thus facilitating the subsequent ligation or condensation. This hypothesis is clearly experimentally testable. If proved, it is of interest to further examine whether Trp RNA-binding attenuation proteins[34] and pentatricopeptide repeat proteins[35], both interacting with RNA in a sequence-specific manner, can serve

a similar purpose as PUF. More importantly, proving these possibilities by rational design of experiments suggests a new protein-based strategy for synthesizing DNA/RNA molecules, which is of interest in life science research and biotechnology.

The potential sequence information transferring from TALE and PUF to DNA and RNA, respectively, however, would not occur in nature, presumably due to certain limitations (e.g., the binding repeats in naturally occurring TALE and PUF proteins were found to be less than 30[13] and 10 (**Table 1**), respectively). In other words, it is unlikely that in nature genetic information is stored in proteins like TALE and PUF, which is then transferred into nucleic acids. In this regard, the hypothesis does not challenge the central dogma.

Nevertheless, the specific amino acid-nucleotide interaction between TALE and DNA, or between PUF and RNA, may provide new insights into the origin of the central dogma. For instance, regarding the origin of the ribosome-mediated translation from mRNA to protein, a direct codon-amino acid stereochemical pairing was suggested to occur during evolution[5,6], which later evolved to the current translation form that does not involve such direct pairing. It was proposed recently[8] that the third characteristic residue of the PUF repeat, which stacks on and sandwiches successive bound RNA bases[19], may have played a role during the evolution of such pairing. Therefore, it is of interest to further investigate the evolutionary significance of the other two characteristic residues of the PUF repeat that make hydrogen bonds or Van der Waals interactions with RNA.**H**

ACKNOWLEDGEMENT This work was supported by research grants from the National Natural Science Foundation of China (No. 31100559 and No. 31270804 to X.F.) and the National Basic Research Program of China (973 Program) (No. 2012CB917300 to X.F.).

ABOUT THE AUTHOR Dr. Fu is a protein scientist focusing on the mechanism of protein biogenesis, folding, assembly and degradation, as well as protein evolution. His long-term goal is to elucidate how the genetic information encoded by DNA is accurately transferred to the three-dimensional structure information of proteins, which is usually facilitated by molecular chaperones and folding catalysts in cells.

REFERENCES

1 Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138-63.
PMid:13580867

2 Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.
http://dx.doi.org/10.1038/227561a0
PMid:4913914

3 McCarthy BJ, Holland JJ. Denatured DNA as a direct template for in vitro protein synthesis. Proc Natl Acad Sci U S A. 1965;54(3):880-6.
http://dx.doi.org/10.1073/pnas.54.3.880
PMid:8900275

4 Hulen C, Legault-Demare J. In vitro synthesis of large peptide molecules using glucosylated single-stranded bacteriophage T4D DNA template. Nucleic Acids Res. 1975;2(11):2037-48.
http://dx.doi.org/10.1093/nar/2.11.2037
PMid:1052527 PMCid:PMC343570

5 Woese CR, Dugre DH, Saxinger WC, Dugre SA. The molecular basis for the genetic code. Proc Natl Acad Sci U S A. 1966;55(4):966-74.
http://dx.doi.org/10.1073/pnas.55.4.966
PMid:5219702 PMCid:PMC224258

6 Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC. On the fundamental nature and evolution of the genetic code. Cold Spring Harb Symp Quant Biol. 1966;31:723-36.
http://dx.doi.org/10.1101/SQB.1966.031.01.093
PMid:5237212

7 Hlevnjak M, Polyansky AA, Zagrovic B. Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. Nucleic Acids Res. 2012;40(18):8874-82.
http://dx.doi.org/10.1093/nar/gks679
PMid:22844092 PMCid:PMC3467073

8 Yarus M, Widmann JJ, Knight R. RNA-amino acid binding: a stereochemical era for the genetic code. J Mol Evol. 2009;69(5):406-29.
http://dx.doi.org/10.1007/s00239-009-9270-1
PMid:19795157

9 Bogdanove AJ, Voytas DF. TAL effectors: customizable proteins for DNA targeting. Science. 2011;333(6051):1843-6.
http://dx.doi.org/10.1126/science.1204094
PMid:21960622

10 Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. Science. 2009;326(5959):1509-12.
http://dx.doi.org/10.1126/science.1178811
PMid:19933107

11 Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. Science. 2009;326(5959):1501.
http://dx.doi.org/10.1126/science.1178817
PMid:19933106

12 Deng D, Yan C, Pan X, Mahfouz M, Wang J, Zhu JK, et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science. 2012;335(6069):720-3.
http://dx.doi.org/10.1126/science.1215670
PMid:22223738 PMCid:PMC3586824

13 Boch J, Bonas U. *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol. 2010;48:419-36.
http://dx.doi.org/10.1146/annurev-phyto-080508-081936
PMid:19400638

14 Saenger W. Principles of Nucleic Acid Structure. New York, Berlin: Springer; 1984.
http://dx.doi.org/10.1007/978-1-4612-5190-3
PMCid:PMC557514

15 Sawada T, Yoshizawa M, Sato S, Fujita M. Minimal nucleotide duplex formation in water through enclathration in self-assembled hosts. Nat Chem. 2009;1(1):53-6.
http://dx.doi.org/10.1038/nchem.100
PMid:21378801

16 Sawada T, Fujita M. A single Watson-Crick G x C base pair in water: aqueous hydrogen bonds in hydrophobic cavities. J Am Chem Soc. 2010;132(20):7194-201.
http://dx.doi.org/10.1021/ja101718c
PMid:20429562

17 Meckler JF, Bhakta MS, Kim MS, Ovadia R, Habrian CH, Zykovich A, et al. Quantitative analysis of TALE-DNA interactions suggests polarity effects. Nucleic Acids Res. 2013;41(7):4118-28.
http://dx.doi.org/10.1093/nar/gkt085
PMid:23408851 PMCid:PMC3627578

18 Wang X, Zamore PD, Hall TM. Crystal structure of a Pumilio homology domain. Mol Cell. 2001;7(4):855-65.
http://dx.doi.org/10.1016/S1097-2765(01)00229-5

19 Wang X, McLachlan J, Zamore PD, Hall TM. Modular recognition of RNA by a human pumilio-homology domain. Cell. 2002;110(4):501-12.
http://dx.doi.org/10.1016/S0092-8674(02)00873-5

20 Filipovska A, Razif MF, Nygard KK, Rackham O. A universal code for RNA recognition by PUF proteins. Nat Chem Biol. 2011;7(7):425-7.
http://dx.doi.org/10.1038/nchembio.577
PMid:21572425

21 Dong S, Wang Y, Cassidy-Amstutz C, Lu G, Bigler R, Jezyk MR, et al. Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. J Biol Chem. 2011;286(30):26732-42.
http://dx.doi.org/10.1074/jbc.M111.244889
PMid:21653694 PMCid:PMC3144504

22 Yin P, Deng D, Yan C, Pan X, Xi JJ, Yan N, et al. Specific DNA-RNA hybrid recognition by TAL effectors. Cell Rep. 2012;2(4):707-13.
http://dx.doi.org/10.1016/j.celrep.2012.09.001
PMid:23022487

23 Sokolova NI, Ashirbekova DT, Dolinnaya NG, Shabarova ZA. Chemical reactions within DNA duplexes. Cyanogen bromide as an effective oligodeoxyribonucleotide coupling agent. FEBS Lett. 1988;232(1):153-5.
http://dx.doi.org/10.1016/0014-5793(88)80406-X

24 Ibanez JD, Kimball AP, Oro J. Possible prebiotic condensation of mononucleotides by cyanamide. Science. 1971;173(3995):444-6.
http://dx.doi.org/10.1126/science.173.3995.444
PMid:17770449

25 Ibanez JD, Kimball AP, Oro J. Condensation of mononucleotides by imidazole. J Mol Evol. 1971;1(1):112-4.
http://dx.doi.org/10.1007/BF01659398
PMid:5173650

26 Grunberg-Manago M, Oritz PJ, Ochoa S. Enzymatic synthesis of nucleic acidlike polynucleotides. Science. 1955;122(3176):907-10.
http://dx.doi.org/10.1126/science.122.3176.907

27 Slomovic S, Portnoy V, Yehudai-Resheff S, Bronshtein E, Schuster G. Polynucleotide phosphorylase and the archaeal exosome as poly(A)-polymerases. Biochim Biophys Acta. 2008;1779(4):247-55.
http://dx.doi.org/10.1016/j.bbagrm.2007.12.004
PMid:18177749

28 England TE, Gumport RI, Uhlenbeck OC. Dinucleoside pyrophosphate are substrates for T4-induced RNA ligase. Proc Natl Acad Sci U S A. 1977;74(11):4839-42.
http://dx.doi.org/10.1073/pnas.74.11.4839
PMid:200936 PMCid:PMC432051

29 England TE, Uhlenbeck OC. Enzymatic oligoribonucleotide synthesis with T4 RNA ligase. Biochemistry. 1978;17(11):2069-76.
http://dx.doi.org/10.1021/bi00604a008

30 Kikuchi Y, Hishinuma F, Sakaguchi K. Addition of mononucleotides to oligoribonucleotide acceptors with T4 RNA ligase. Proc Natl Acad Sci U S A. 1978;75(3):1270-3.
http://dx.doi.org/10.1073/pnas.75.3.1270
PMid:274717 PMCid:PMC411452

31 Kanaya E, Yanagawa H. Template-directed polymerization of oligoadenylates using cyanogen bromide. Biochemistry. 1986;25(23):7423-30.
http://dx.doi.org/10.1021/bi00371a026

32 Kawamura K, Ferris JP. Kinetic and mechanistic analysis of dinucleotide and oligonucleotide formation from the 5'-phosphorimidazolide of adenosine on Na(+)-montmorillonite. J Am Chem Soc. 1994;116(17):7564-72.
http://dx.doi.org/10.1021/ja00096a013

33 Ferris JP, Ertem G. Oligomerization of ribonucleotides on montmorillonite: reaction of the 5'-phosphorimidazolide of adenosine. Science. 1992;257(5075):1387-9.
http://dx.doi.org/10.1126/science.1529338

34 Antson AA, Dodson EJ, Dodson G, Greaves RB, Chen X, Gollnick P. Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. Nature. 1999;401(6750):235-42.
http://dx.doi.org/10.1038/45730
PMid:10499579

35 Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008;13(12):663-70.
http://dx.doi.org/10.1016/j.tplants.2008.10.001
PMid:19004664